Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.201744400
Vol. 4, No. 4, Desember 2017, hlm. 263-267 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 51/E/KPT/2017 e-ISSN: 2528-6579

OPTIMASI SUFFIX TREE CLUSTERING DENGAN WORDNET DAN NAMED ENTITY RECOGNITION UNTUK PENGELOMPOKAN DOKUMEN

Satrio Hadi Wijoyo[1], Admaja Dwi Herlambang[2], Fahrur Rozi[3], Septiyan Andika Isanta[4]

[1,2]Jurusan Sistem Informasi, Fakultas Ilmu Komputer, Universitas Brawijaya
[3]Jurusan Pendidikan Teknologi Informasi, STKIP PGRI Tulungagung
[4]Jurusan Teknik Informatika, Universitas Muhammadiyah Malang
Email: [1]satriohadi@ub.ac.id, [2]herlambang@ub.ac.id, [3]rozifahrur04@gmail.com,
[4]septiyan.andika@gmail.com

(Naskah masuk: 26 Juli 2017, diterima untuk diterbitkan: 17 Desember 2017)

Abstrak

Semakin meningkatnya jumlah dokumen teks di dunia digital mempengaruhi banyaknya jumlah informasi dan menyebabkan kesulitan dalam proses temu kembali informasi (information retreival). Clustering dokumen merupakan suatu bidang text mining yang penting dan dapat digunakan untuk mengefisienkan dalam pengelolaan teks serta peringkasan teks. Namun beberapa permasalahan muncul dalam clustering dokumen teks terutama dalam dokumen berita seperti ambiguitas dalam content, overlapping cluster, dan struktur unik yang terdapat dalam dokumen berita. Penelitian ini mengusulkan metode baru yaitu optimasi Suffix Tree Clustering (STC) dengan WordNet dan Named Entity Recognition (NER) untuk pengelompokan dokumen. Metode ini memiliki beberapa tahap, yaitu prepocessing dokumen dengan mengekstraksi named entity serta melakukan deteksi sinonim berdasarkan WordNet. Tahap kedua adalah pembobotan term dengan tfidf dan nerfidf. Tahap ketiga adalah melakukan clustering dokumen dengan menggunakan Suffix Tree Clustering. Berdasarkan pengujian didapatkan rata-rata nilai precision sebesar 79.83%, recall 77.25%, dan f-measure 78.30 %.

Kata kunci: *Clustering dokumen, Named Entity Recognition, Suffix Tree Clustering, WordNet*

Abstract

The increasingnumber oftext documentsin the internet, influence on the number of information and lead to difficulties in the process of information retrieval. Documents clustering is main field of text mining and can be used to stream line the management of text and summarization of text. However, some problems a risein documents clustering, especially in news documents such as ambiguity in the content, overlapping clusters, and theuniquestructure ofthe news thatcontained inthe document. Inthisresearch, we proposea newmethodfor documents clustering, optimization Suffix Tree Clustering (STC) with WordNet and Named Entity Recognition (NER). In this method there are several step, step one is prepocessing documents with named entity extraction and synonym detection based on WordNet. Step two is term weighting with tfidf and nerfidf. For the last step is document clustering using Suffix Tree Clustering. Based on testingwe obtained 79.83% for precision, 77.25% for recall, and 78.30% for F-measure

Keywords: *Documents Clustering, Named Entity Recognition, Suffix Tree Clustering, WordNet*

1. PENDAHULUAN

Pertumbuhan dunia digital yang pesat terutama di *World Wide Web* menyebabkan meningkatnya volume dokumen teks secara besar-besaran. Meningkatnya volume dokumen teks ini berpengaruh terhadap jumlah informasi yang sangat banyak dan menyebabkan kesulitan dalam proses temu kembali informasi (*information retreival*). Sehingga dibutuhkan suatu metode yang mampu mengorganisasikan dokumen teks ke dalam informasi yang mudah dipahami oleh pengguna serta dalam meningkatkan efisiensi dalam *information retreival* (Nogueira et al, 2011).

Clustering dokumen merupakan suatu bidang *text mining* yang penting dan dapat digunakan untuk mengefisienkan dalam pengelolaan teks serta peringkasan teks (Luo et al, 2009). *Clustering* dalam suatu dokumen dapat membantu mengelompokkan dokumen berdasarkan *content* yang tepat, sehingga dapat membantu pangguna mendapat informasi yang diinginkan secara tepat.

Namun, terdapat beberapa permasalahan dalam *clustering* dokumen. Selain permasalahan terhadap volume dokumen yang mempengaruhi skalabilitas, permasalahan mengenai *content* juga berpengaruh dalam *clustering* dokumen. Contoh, dalam artikel berita, terkadang beberapa artikel dikategorikan dalam kategori yang sama padahal tidak memiliki kata-kata yang mirip. Begitu juga sebaliknya suatu artikel terkadang dikategorikan dalam kategori yang berbeda padahal memiliki kata-kata yang mirip (Bouras et al, 2012).

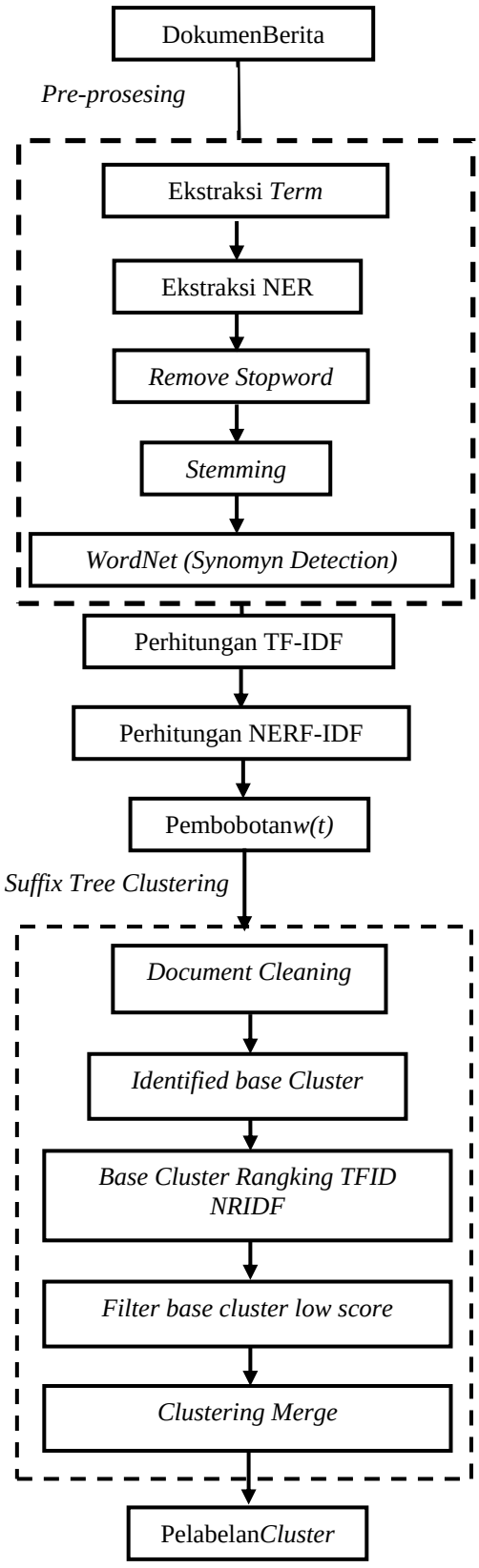
Penelitian dari Bouras pada tahun 2012,

menggunakan metode *k-means* yang terintegrasi dengan *WordNet* untuk mendeteksi hubungan semantik yang terjadi antar *term* untuk dokumen berita. Dengan memperhatikan semantik, permasalahan ambiguitas *content* dalam *clustering* dapat terselesaikan. Penggunaan metode k-means memiliki keterbatasan dalam *clustering* dokumen, karena algoritma *k-means* memperlakukan suatu dokumen sebagai kumpulan kata-kata dan mengabaikan *sequence* kata dalam dokumen serta *k-means* memerlukan suatu *stop-condition* dan nilai *k* sebagai *initial* awal masukan jumlah *cluster*. Selain itu *k-means* memiliki masalah utama yaitu *overlapping cluster* yang hanya dapat menempatkan suatu dokumen tepat pada satu *cluster*, padahal suatu dokumen dimungkinkan menempati lebih dari satu *cluster*. Selain itu terdapat struktur yang unik dalam dokumen berita, yaitu dalam dokumen berita sebagian besar tersusun atas struktur "*time*", "*location*", "*character*" dan "*event*". Dengan mengekstraksi struktur tersebut akan didapatkan sebuah *noun* ataupun *meaningful* frase yang dapat digunakan sebagai *cluster* label (Zhang et al, 2013). Sehingga permasalahan yang terdapat dalam *k-means* terutama *overlapping cluster* serta ekstraksi struktur dalam dokumen berita menjadi tantangan baru dalam *clustering* dokumen.

Penelitian ini bertujuan untuk mengembangkan metode baru yaitu optimasi *Suffix Tree Clustering* (STC) dengan *WordNet* dan *Named Entity Recognition* (NER) untuk pengelompokan dokumen. Penggunaan algoritma STC tepat untuk mengatasi kelemahan pada *k-means* yang terdapat dalam penelitian dari Bouras menangani *overlapping clustering*. Hal ini dikarenakan STC memperlakukan kata-kata hasil ekstraksi sebagai suatu koleksi kata-kata yang memiliki hubungan terhadap suatu dokumen. Serta STC menggunakan salah satu frase atau kata sebagai topik utama atau label *cluster*. Selain itu penggunaan algoritma NER dalam penelitian ini dapat mengekstraksi struktur yang terdapat dalam dokumen berita (Zhang et al, 2013).

2. METODE

Perancangan sistem temu kembali informasi yang dibangun dalam penelitian ini adalah sistem untuk pengelompokan dokumen berita yang mempertimbangkan *similarity* kata dan *meta-data* dalam dokumen tersebut. Metode optimasi STC dengan *WordNet* dan NER memiliki beberapa tahap, yaitu *prepocessing* dokumen dengan mengekstraksi *named entity* serta melakukan deteksi sinonim berdasarkan *WordNet*. Tahap kedua adalah pembobotan *term* dengan tfidf dan nerfidf. Tahap ketiga adalah melakukan *clustering* dokumen dengan menggunakan STC. Detail tentang metode penelitian yang diusulkan digambarkan pada Gambar 1.



Gambar 1. Alur Kerja Metode yang Diusulkan

Pada tahap *prepocessing* dokumen dilakukan ekstraksi *term* terlebih dahulu. Ekstraksi *term* digunakan untuk mengekstrasi ciri-ciri dari suatu koleksi dokumen berita yang sering disebut himpunan *term*. Himpunan *term* yang bermakna umum dilakukan *remove stopword* dengan menggunakan daftar *stopword* untuk bahasa Inggris. Kemudian dikembalikan kebentuk kata dasarnya dengan menggunakan algoritma *stemming* serta dapat meningktakan performa IR (*Information Retrieval*). Salah satu cara mentransformasi kata yang berimbuan dalam dokumen ke bentuk kata dasarnya disebut Algoritma *stemming*.

Setelah didapatkan *term* setiap dokumen akan dilakukan ekstraksi NER. Ekstraksi NER dilakukan untuk menemukan dan mengenali entitas nama (nama orang, nama organisasi, dan nama lokasi), ekspresi waktu (tanggal, jam, dan durasi) dan ekstraksi angka (uang, persentasi, ukuran, dan kardinal) dari dokumen berita. Setelah proses NER dijalankan, akan didapatkan *named-entity* (NE) atau sering disebut *mention* (fitur nama orang terkait) beserta tipe entitasnya, seperti kata "Amir" sebagai entitas nama orang, kata "13.00" sebagai entitas waktu, dan kata "Beijing" sebagai entitas lokasi. Deteksi NE dapat dilakukan dengan melihat pola dari kalimat yang ada di dalam dokumen berita.

Seluruh *term* yang didapatkan akan dilakukan pendeteksian sinonim berdasarkan *WordNet*. *WordNet* merupakan sistem *lexical database* yang menyimpan informasi relasi semantik antar *synset* (*Synonym set*). Makna sama yang dapat saling menggantikan dalam konteks tertentu yang dimiliki kumpulan satu kata atau lebih disebut *synset.*

Tahap pembobotan *term* merupakan perhitungan frekuensi kemunculan *term* (kata) dalam dokumen berita serta pembobotannya. Perhitungan pembobotan *term* ($tfidf$) didapatkan melalui Persamaan (1)

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D), \qquad (1)$$

dimana $tfidf(t,d,D)$ adalah frekuensi *term* dari dokumen berita d, D adalah total dokumen berita, t adalah banyaknya *term*, $tf(t,d)$ adalah *term frequency* dari suatu dokumen berita, dan $idf(t,D)$adalah *inverse document frequency* (Luo et al, 2009). Setelah ditemukan nilai $tfidf(t,d,D)$, selanjutnya nilai $tfidf(t,d,D)$ dilakukan normalisasi yang ditunjukkan dalam persamaan (2)

$$tfidf(t,d)_{norm} = \frac{tfidf(t,d,D)}{\max(tfidf(t,d,D))}, \qquad (2)$$

dimana $tfidf(t,d)_{norm}$ merupakan nilai hasil normalisasi, dan $\max(tfidf(t,d,D))$ merupakan nilai maksimum $tfidf$ dalam dokumen d. Selanjutnya adalah pembobotan terhadap perhitungan kemunculan entitas-entitas dalam dokumen berita. Perhitungan frekuensi entitas didapatkan melalui Persamaan (3)

$$nerfidf(ner,d,D) = nerf(ner,d) \times idf(ner,D), \qquad (3)$$

dimana $nerfidf(ner,d,D)$ adalah frekuensi entitas dari dokumen berita d, D adalah total dokumen berita neradalah banyaknya entitas di dokumen, $nerf(ner,d)$adalah NER *frequency* dari suatu dokumen berita, dan $idf(ner,D)$adalah *inverse document frequency*. Selanjutnya nilai $nerfidf(ner,d)$ dilakukan normalisasi melalui Persamaan (4)

$$nerfidf(ner,d)_{norm} = \frac{nerfidf(ner,d,D)+1}{\max(nerfidf(ner,d,D))}, \qquad (4)$$

dimana $nerfidf(ner,d)_{norm}$ merupakan nilai normalisasi dari $nerfidf(ner,d)$, dan $\max(nerfidf(d))$ merupakan nilai maksimum $nerfidf$ dalam dokumen d. Penambahan angka 1 terhadap $nerfidf(ner,d)$ dilakukan agar $nerfidf(ner,d)_{norm}$ tidak bernilai 0.

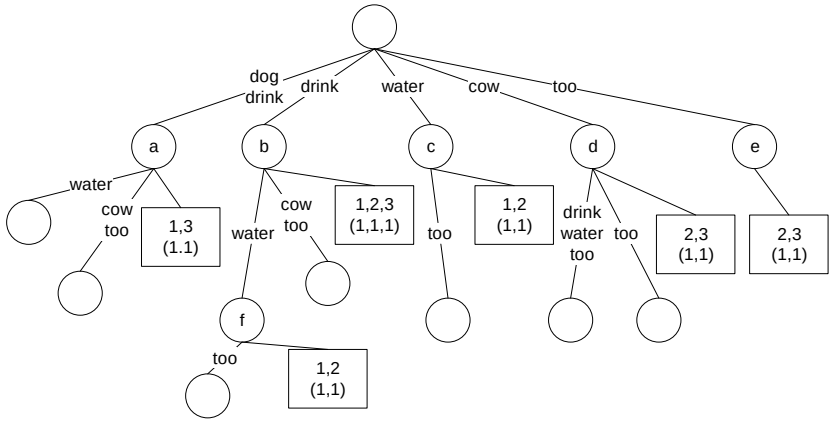
Setelah ditemukan nilai $tfidf(t,d)_{norm}$ dan $nerfidf(ner,d)_{norm}$untuk masing – masing *term* maka selanjutnya dilakukan perhitungan pembobotan *term* dokumen berita sebelum dilakukan proses pengelompokan. Pada penelitian ini mengusulkan metode pembobotan didapatkan dari kombinasi perhitungan TF-IDF dan NERF-IDF melalui Persamaan (5).

$$w(t) = (tfidf_{norm}) * (nerfidf_{norm}), \qquad (5)$$

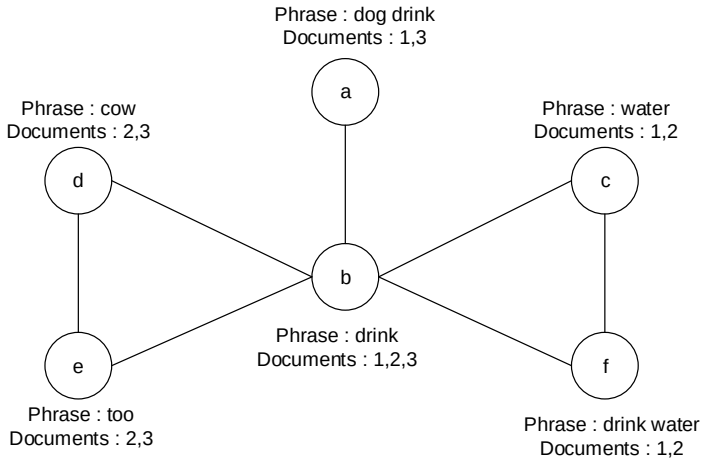
dimana$w(t)$ adalah bobot dari *term* dan NER dari dokumen berita.

Tahap pengelompokan menggunakan algoritma *suffix tree clustering*. Terdapat dua tahapan utama dalam algoritma *Suffix tree clustering* untuk proses pengelompokan. Tahapan pertama adalah pencarian *shared phrase* pada semua koleksi dokumen berita dan disebut juga sebagai *phrase cluster* atau *base cluster.* Tahapan kedua adalah kombinasi *base cluster-base cluster* ke dalam suatu *cluster*. Kombinasi antar dua *base cluster* berdasarkan jumlah dokumen berita yang terdapat *overlap* diantara kedua *base cluster* tersebut seperti pada Gambar 2 dan 3 (Worawitphinyo, 2011).

Pelabelan *cluster* dalam *suffix tree clustering* diperoleh dari *shared pharse* yang ada dalam masing-masing *cluster*. Dikarenakan *shared pharse* bisa berisi banyak pharse, di paper ini dilakukan pemfilteran terhadap setengah bagian teratas yang telah di *ranking* berdasarakan bobot $tfidf$ dan $nerfidf$.



Gambar 2. *Suffix Tree Clustering* dari *Tree* Dokumen



Gambar 3. Grafik dari *Base Cluster*

3. UJI COBA DAN PEMBAHASAN

Penelitian mengenai dokumen clustering ini menggunakan kumpulan data dari artikel berita yang terdiri atas 300 artikel yang bersumber dari 20 Newsgroups yang diperoleh dari situs: http://web.istl.utl.pt/~acardoso/datasets. Data tersebut digunakan untuk single label, dalam paper ini dilakukan modifikasi dari data single label menjadi multilabel jadi data digroupkan lagi.

Penelitian ini dilakukan 2 percobaan, percobaan pertama hanya menggunakan STC, percobaan kedua menggunakan STC dengan *WordNet* serta NER. Pengujian efektifitas dari metode *clustering* menggunakan *precision* (*P*), *recall* (*R*), dan *f-measure*. Dokumen terpanggil yang relevan dengan pernyataan (*query*) yang dimasukkan pengguna dalam suatu sistem temu balik informasi disebut *recall*. Sedangkan kemampuan sistem menemukan jumlah kelompok dokumen relevan dari total jumlah dokumen disebut *precision*. Perhitungan kombinasi antara *recall* dan *precision* disebut *F-measure* (Tan, 2006).

Berdasarkan penelitian yang dilakukan oleh Worawitphinyo pada tahun 2011, *filtering base cluster low score* menggunakan *threshold* 50% untuk menyeleksi *base cluster* teratas dengan menggunakan nilai $tfidf$. Sementara pada penelitian ini menggunakan nilai $w(t)$ sebagai pengganti $tfidf$ untuk menyeleksi *base cluster*.

Tabel 1 menunjukkan bahwa penggunaan *WordNet* dan NER dalam STC dapat meningkatkan nilai *precission* sebuah *cluster*. Semua *cluster* mengalami peningkatan yang berbeda-beda nilainya. Peningkatan nilai *precission* yang signifikan terdapat pada *cluster* no 5 dimana menggunakan metode STC dengan nilai *precission* sebesar 62.5%. Percobaan keduasetelah menggunakan STC, *WordNet*, dan NER nilai *precission* meningkat menjadi 70%.

Tabel 1. Berbandingan Nilai *Precission* dari Kedua Percobaan

Cluster	**STC**	**STC *WORDNET* dan NER**
1	77.8	80
2	87.5	88.9
3	69.2	75
4	76.9	83.3
5	62.5	70

6	77.8	81.8

Tabel 2. Berbandingan Nilai *Recall* dari Kedua Percobaan

Cluster	STC	STC *WORDNET* dan NER
1	63.6	72.7
2	63.6	72.7
3	81.8	81.8
4	90.9	90.9
5	54.5	63.6
6	63.6	81.8

Tabel 3. Berbandingan Nilai *F-Meausre* dari Kedua Percobaan

Cluster	STC	STC *WORDNET* dan NER
1	69.9	76.1
2	73.7	79.9
3	74.9	78.3
4	83.3	86.9
5	58.2	66.6
6	69.9	81.8

Tabel 2 menunjukkan bahwa penggunaan *WordNet* dan NER dalam STC dapat meningkatkan nilai *recall* sebuah *cluster*. Semua *cluster* mengalami peningkatan nilai *recall* yang berbeda-beda. Nilai peningkatan *recall* yang signifikan terdapat pada *cluster* no 6 dimana menggunakan metode STC dengan nilai *precission* sebesar 63.6%. Percobaan kedua setelah menggunakan STC, *WordNet*, dan NER nilai *precission* meningkat menjadi 81.8%.

Tabel 3 menunjukkan bahwa penggunaan *WordNet* dan NER dalam STC dapat meningkatkan nilai *f-measure* sebuah *cluster*. Semua cluster mengalami peningkatan nilai *recall* yang berbeda-beda. Nilai peningkatan *recall* yang signikan terdapat pada *cluster* no 6 dimana menggunakan metode STC dengan nilai *f-meausre* sebesar 69.9%. Percobaan kedua setelah menggunakan STC, *WordNet,* dan NER nilai *f-meausre* meningkatsebesar 81.8%.

WordNet dan NER dapat meningkatkan nilai *percission, recall, dan f-measure* STC karena STC berdasarkan *sharing pharse*. *WordNet* (*synonym detection*) dapat mendeteksi kata yang beda penulisan tapi sama makna. Sehingga *phrase* atau kata yang sama antar dokumen bertambah banyak. NER sendiri berguna untuk mendeteksi dokumen yang mempunyai entitas. NER juga dapat mendeteksi kata yang dianggap penting.

Synonym detection telah mampu membuktikan meningkatkan kualitas *clustering* dokumen lebih baik. Namun, peningkatan kualitas *clustering* dapat lebih baik lagi jika dapat ditambahkan *hyponym* dan *hypernym detection* yang terintegrasi dengan *WordNet*.Sehingga untuk pengembangan penelitian selanjutnya dapat ditambahkan *hyponym* dan *hypernym detection* dalam *clustering* dokumen.

4. KESIMPULAN

Hasil percobaan dari penelitian ini menunjukkan bahwa metode yang diusulkan dapat melakukan pengelompokan dokumen dengan sangat baik. *WordNet* (*synonym detction*) dapat mendeteksi kata yang beda penulisan tapi sama makna. NER dapat mendeteksi dokumen yang mempunyai entitas. Selain itu, *WordNet* dan NER dapat digunakan untuk optimasi *clustering* dokumen menggunakan STC.

5. DAFTAR PUSTAKA

NOGUEIRA, T. M., CAMARGO, H. A., & REZENDE, S. O. 2011. Fuzzy Rules for Document Classification to Improve Information Retrieval.International Journal of Computer Information Systems and Industrial Management Applications, 3, 210-217.

LUO, CONGNAN, LI, YANJUN, CHUNG, SOON M. 2009. Text document clustering based on neighbors. Data & Knowledge Engineering. 1271-1288.

BOURAS, CHRISTOS, TSOGKAS, VASILIS, 2012. A Clustering Technique for News Articles using *WordNet*. Knowledge-Based Systems. 115-128.

ZHANG, J.,DANG, Q., LU, Y., SUN, S., 2013. Suffix Tree Clustering with Named Entity Recognition. International Conference on Cloud Computing and Big Data. 549-556.

WORAWITPHINYO PHIRADIT, GAO XIAOYING, JABEEN SHAHIDA, 2011. Improving Suffix Tree Clustering with New Ranking and Similarity Measures. 7th International Conference, ADMA 2011.

TAN, P. N., MICHEAL S., & VIPIN K. 2006.Introduction to Data Mining. Pearson Education : India.