Text Mining Dan Pola Algoritma Dalam Penyelesaian Masalah Informasi : (Sebuah Ulasan)

Ali Firdaus*[1] ,Wahyu Istalama Firdaus*[2]
[1*] Program Studi Teknologi Informatika Multimedia Digital, Politeknik Negeri Sriwijaya,
[2*] Faculty of Islamic Studies, International University of Aprica
[1*]alifirdaus1970@gmail.com, [2*] wahyuistalama02@gmail.com

Abstrak

Text Mining bertujuan untuk menemukan informasi berharga yang tersembunyi baik dari sumber informasi terstruktur dan tidak terstruktur. Web merupakan sumber utama tempat keberadaan text yang menyimpan informasi tekstual yang tersedia bagi kita. Jumlah text ini seiring waktu terus mengalami peningkatan secara terus menerus. Text Mining merupakan suatu penemuan baru yang sebelumnya informasinya tidak diketahui. Informasi yang diekstrak dari berbagai sumber daya tertulis dilakukan secara otomatis. Elemen kuncinya adalah menghubungkan beberapa informasi yang diekstraksi menjadi satu sehingga dapat membentuk fakta baru atau hipotesis baru untuk dieksplorasi lebih lanjut. Pada makalah ini hanya fokus untuk meninjau dan memberikan ulasan tentang pola dan konsep dasar dari berbagai teknik text mining yang banyak digunakan untuk penyelesaian masalah informasi yang sudah dipulikasi oleh beberapa orang penulis. Selain itu penulis juga menunjukkan ulasan utama tentang teks mining dan tema utamanya pada sekitar tahun 1980 hingga 2010-an. yang semula berawal dari ilmu informasi menjadi sistem informasi serta merambah ke manajemen teknologi, di bidang-bidang arsitektur, dan ekologi sosial.

Kata Kunci : Teks Mining, Tren, Pola, Algoritma, Informasi

Abstract

Text Mining to find aims valuable hidden information from both structured and unstructured sources of information. The web is the main source of textual existence which stores the textual information available to us. The amount of text is continuously increasing over time. Text Mining is a new discovery whose information was previously unknown. Information extracted from various written resources is done automatically. The key element is to connect some of the extracted information into one so that it can form new facts or new hypotheses for further exploration. This paper only focuses on reviewing and providing an overview of the basic patterns and concepts of various text mining techniques that are widely used for solving information problems that have been published by several authors. In addition, the author also shows a major review of text mining and its main themes from around the 1980s to 2010s. which originally started from information science into information systems and penetrated into technology management, in the fields of architecture, and social ecology.

Keywords : Text Mining, Trends, Patterns, Algorithms, Information

1. PENDAHULUAN

Perkembangan teknologi informasi yang semakin pesat serta dukungan jaringan komputer dan komunikasi yang sangat luas (Internet) perlahan lahan telah menjadi bagian yang menyatu dan sulit untuk pisahkan dari diri manusia karena telah menjadi suatu

kebutuhan dan gaya dalam kehidupan masyarakat. Media web serta media jaringan sosial akan terus menghasilkan data *text* yang banyak dan tidak terstruktur, seperti blog pribadi atau kelompok dan perusahaan, facebook, forum atau grup, postingan bebas, dokumentasi, dan lain sebagainya.

Ini menunjukkan perilaku orang dan berpikir secara intuitif (mempunyai kemampuan memahami sesuatu tanpa melalui penalaran rasional dan intelektualitas). Disini banyak mengandung sebaran informasi yang sangat sulit untuk ditangani karena jumlahnya yang besar serta dalam format file dan ukuran yang berbeda beda. Namun disisilain terjadi peningkatan permintaan untuk melakukan analisis data teks. Oleh karena itu berlandaskan kepada masalah-masalah yang terjadi dan alasan yang telah diuraikan di atas maka bagaimana cara untuk memperoleh informasi yang benar-benar dibutuhkan oleh masyarakat umum dari sejumlah besar data teks yang tidak terstruktur menjadi topik utama didalam riset para peneliti di bidang data mining dan informasi. Dan karena alasan itu pula Pada penelitian ini penulis hanya fokus untuk meneliti dan memberikan ulasan tentang pola dan konsep dasar dari berbagai teknik *text mining* yang banyak digunakan para peneliti untuk penyelesaian masalah informasi yang terkait dengan *text mining* yang sudah dipublikasikan.

Text Mining merupakan penemuan pengetahuan di database dalam bentuk tekstual (knowledge discovery in textual database atau disingkat dengan KDT)[65][17], bisa disebut juga penggalian atau pencarian data-data yang berbentuk teks[32] yang di antaranya adalah ketertarikan terhadap pengetahuan yang baru dibuat, diartikan sebagai bagian dari proses penggalian atau pencarian data *text* yang sebelumnya tidak diketahui, sehingga dapat dimengerti, mempunyai potensi dan pola praktis atau pengetahuan dari koleksi data teks atau corpus masif (kumpulan teks yang menangkap penggunaan bahasa dalam bentuk tertulis atau lisan secara utuh dan padat) dan tidak terstruktur.

Text mining sebagai ilmu pengetahuan cabang dari data mining, dipercaya memiliki nilai komersial yang jau lebih tinggi dibandingkan data mining itu sendiri, karena 80% pada setiap perusahaan terdapat dokumen informasi dalam bentuk teks [62]. Namun, penggalian dan pencarian data teks lebih kompleks dengan pola *text* tidak terstruktur selalu menyulitkan. Text mining merupakan lingkup penelitian yang komprehensif, yang masuk hampir disetiap lini kehidupan kita. Pada makalah ini akan dijelaskan tentang sedikit sejara text mining serta tujuan dan maksud penelitian. Dan akan dijelaskan juga beberapa pola atau model secara umum yang, serta melakukan pengelompokan masalah-masalah yang sering muncul pada text mining. Selain itu penulis juga berusaha membuat ringkasan dalam ulasan ini.

2. METODE PENELITIAN

Sistem intelijen bisnis dapat mewujudkan ekstraksi dan pengkodean dokumen secara otomatis dengan menggunakan komputer, dan menerapkan pengelompokan dan klasifikasi dokumen melalui frekuensi kata secara statistik[34]. Inilah yang menjadi defenisi awal dari sebuah *Business Intelligence*, dan selanjutnya ini dikembangkan sebagai *prototipe text mining*. Setelah itu banyak para pakar dan peneliti yang menggali lebih dalam hingga pekerjaan mereka dibidang penelitian ini membuahkan hasil. Melalui sebuah artikel yang diterbitkan oleh Maron ditahun 1960, dia mempublikasikan cara atau literatur teknik terbaru untuk melakukan pengindeksasian *text* dan sistem mekanisme pencarian pada perpustakaan[46]. Pengelompokan dan klasifikasi secara otomatis (knowledge discovery in textual database) atau KDT diusulkan pertama kali oleh F. Ronen pada Konferensi Internasional Penemuan Pengetahuan dan Penambangan Data yang pertama pada tahun 1995.[17]

Didalam tulisannya Bjornar L. dan Chinatsu A. telah memberikan penjelasan tentang sistem pengelompokan teks dalam rentang waktu yang hampir bersamaan secara mandiri, yang cepat dan efektif, dan terdapat beberapa pilihan algoritma pada setiap fasenya.[38]. Untuk

melihat capaian kualitas hierarki yang dihasilkan mereka menggunakan F-Measure (kombinasi presisi dan pemanggilan ulang). Masih banyak hasil penelitian lainnya yang sangat luar biasa, seperti penelitian yang membahas tentang pengelompokan dan klasifikasi algoritma text mining.[66]-[33], serta aplikasi text mining diberbagai bidang, seperti literatur pertambangan dalam biologi molekuler.[12]-[67], Analisis sentimen atau opinion mining merupakan metode analisis berbasis komputasi mengenai pendapat, sentimen, dan emosi untuk melihat ke-cenderungan suatu sentimen atau pendapat, apakah pendapat tersebut cenderung beropini positif atau negatif.[41].

Penelitian dimensi reduksi data dalam fitur ekstraksi.[36]-[25], representasi teks.[63] dan konstruksi model.[65][64]-[23], pengelompokan dan klasifikasi teks.[26]-[42], kedalaman semantic mining berdasarkan proses bahasa alami.[6]-[10], prediksi nilai saham pada bidang keuangan dan sekuritas.[45], web mining secara global di internet. [8]-[22], perpustakaan digital.[69], Hemalatha, Varma, & Govardhan, pada tahun 2012 melakukan penelitian Stemming yaitu sebuah proses untuk mendapatkan kata dasar dengan menghilangkan imbuhan pada kata. Pada penelitian ini, proses stemmer menggunakan algoritma 3 Confix-Stripping Stemmer yang merupakan pengembangan dari algoritma Nazief and Adriani's Stemmer.[27].

Penelitian lainnya yaiu penggunaan Shared Nearest Neighbor(SNN) untuk pengelompokkan data tiga dimensi clud computing. Penelitian penggunaan SNN dalam pengelompokkan spectral dilakukan oleh He et al.[28]. Penelitian tentang penggunaan SNN juga pernah dilakukan oleh Zahrotun, Dalam penelitian ini membandingkan dua metode similarity yaitu cosine similarity dan jaccard similarity[72]. Penelitian tentang kedekatan atau sering disebut sebagai similarity pernah dilakukan sebelumnya oleh Patidar et al., dalam penelitian ini membandingkan beberapa similarity dan dari beberapa similarity sebelumnya untuk menghasilkan Euclidean similarity yang memiliki akurasi paling baik.[55]. Shared Nearest Neighbor merupakan cara yang terbaik untuk mengelompokkan data dalam jumlah besar. Dengan mengunakan teknik similarity atau kesamaan, setelah ketetanggaan terdekat dari semua titik data telah ditentukan, maka nilai kesamaan yang baru diantara titik-titik data ditentukan dari jumlah ketetanggaan yang dimiliki secara bersama-sama.[68], Sesungguhnya *Text mining* sekarang ini telah menjadi bentuk aplikasi yang praktis seperti Text Miner [30], VisualText [31], IBM Intelligent Miner for Text [29].

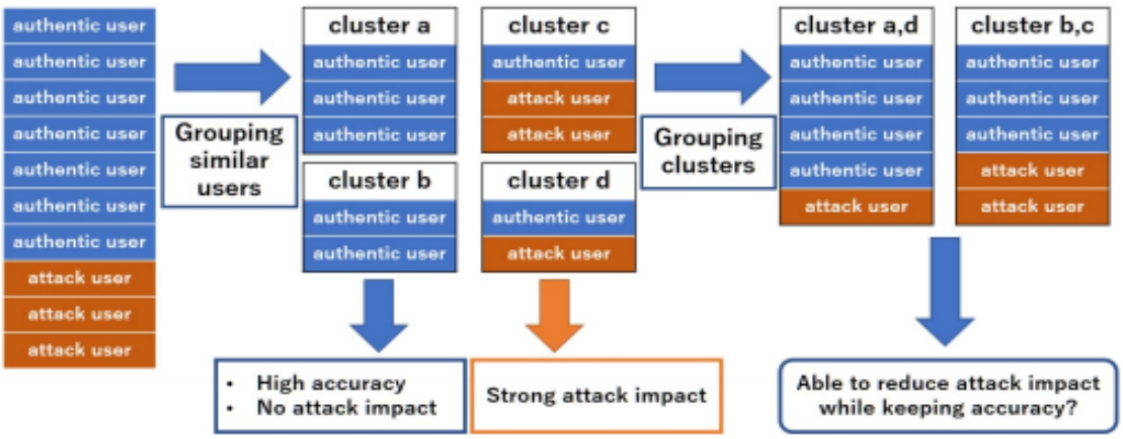
3. HASIL DAN PEMBAHASAN

Pada Text mining selalu melibatkan pra proses dokumen yaitu ; melakukan kategorisasi teks, melakukan ekstraksi informasi, dan mengekstraksi kata. Metode ini untuk mengekstraksi informasi yang diambil dari sumber data dengan cara mengidentifikasi serta melakukan eksplorasi pola yang menarik [20]. Text Mining merupakan teknik yang digunakan untuk menangani permasalahan klasifikasi, clustering, information extraction dan information retrival [4]. Secara Umum Text mining terdiri dari tiga langkah yaitu: teks preprocessing, operasi penggalian teks, postprocessing. Tugas Teks preprocessing adalah termasuk di dalamnya pemilihan data, klasifikasi dan ekstraksi fitur untuk mengubah dokumen menjadi bentuk perantara, yang harus cocok dengan tujuan pencarian yang berbeda. Bagian utama dari Pekerjaan Operasi teks mining mencakup pengelompokan, penemuan aturan asosiasi, tren analisis , pola penemuan, serta algoritma penemuan pengetahuan. Pekerjaan selanjutnya sebelum pemrosesan memanipulasi data atau informasi terbaru yang berasal dari prose teks mining, seperti evaluasi dan pemilihan informasi yang ditemukan, interpretasi dan visualisasi informasi yang dihasilkan.

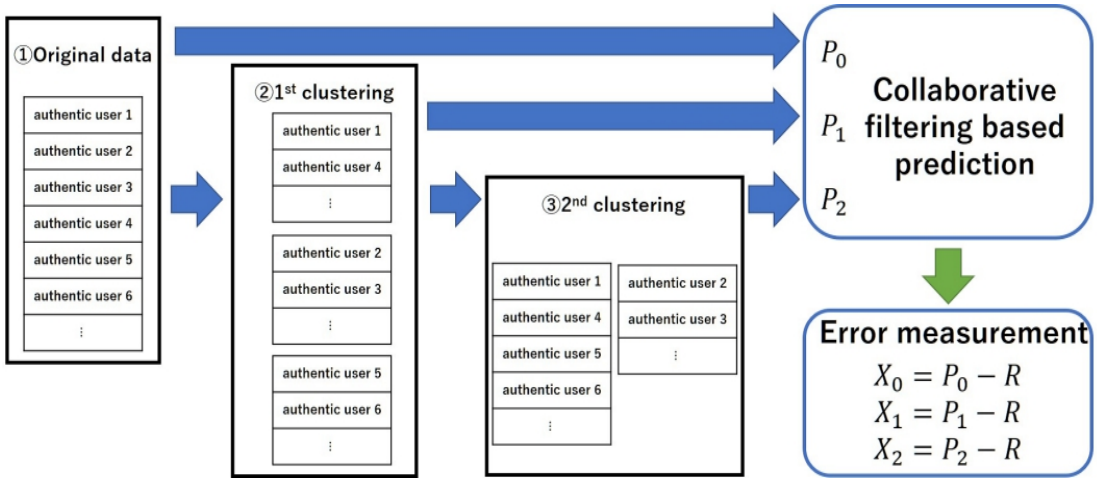
Sampai saat ini, sudah banyak model atau pola penggalian text mining yang telah digunakan. Tren model yang digunakan saat ini adalah Robust Collaborative Filtering yang diusulkan oleh R. Burke, M.P. O'Mahony, N.J. Hurley[58] dan model ini kembali diangkat oleh

"Jianwei Zhang, *Faculty of Science and Engineering Iwate University* Morioka, Japan " dalam tulisannya "Robust Collaborative Filtering Based on Multiple Clustering" yang dipublikasikan pada "2019 IEEE 7th International Conference on Computer Science And Network Technology (ICCSNT) ©2019 IEEE Dalian, China•Oct 19-20, 2019".

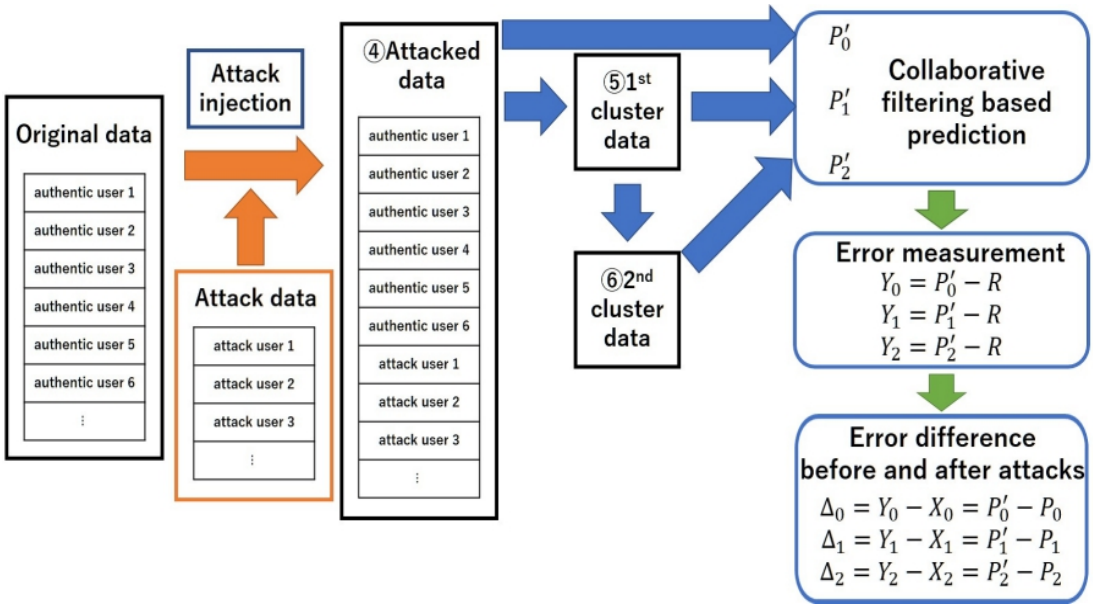
Dalam tulisannya Zhang memecahkan permasalahan yang berhubungan dengan adanya serangan yang bermaksud untuk mendistorsi peringkat yang diprediksi dari item tertentu. Dimana Pemfilteran kolaboratif ini banyak digunakan oleh vendor online dan situs yang memberikan ulasan untuk merekomendasikan peringkat item berdasarkan banyaknya pengguna. Zhang mengusulkan metode pemfilteran kolaboratif yang mengurangi dampak serangan sambil mempertahankan atau meningkatkan akurasi prediksi dengan menerapkan pengelompokan ke target data beberapa kali dan memprediksi peringkat untuk item yang tidak berperingkat di dalam setiap kluster. Pada sisi yang lain penggunaan metode ini amati dengan menggunakan metode evaluasi guna mengukur kesalahan antara peringkat pengguna aktual dan peringkat yang diprediksi. Selain itu, ketahanan terhadap serangan diselidiki dengan membandingkan kesalahan prediksi sebelum dan sesudah serangan. Berikut ini adalah cara prediksi pada sistem Pemfilteran kolaboratif.



Gambar 1 Proses Prediksi Pada Pemfilteran kolaboratif

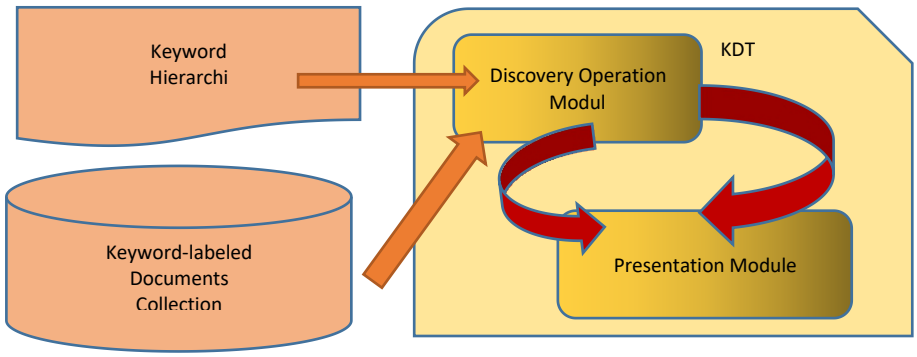


Gambar 2 Perhitungan Prediksi rekomendasi peringkat berbasis Pemfilteran kolaboratif



Gbr. 3. Pengukuran kesalahan sebelum serangan Pada Pemfilteran kolaboratif

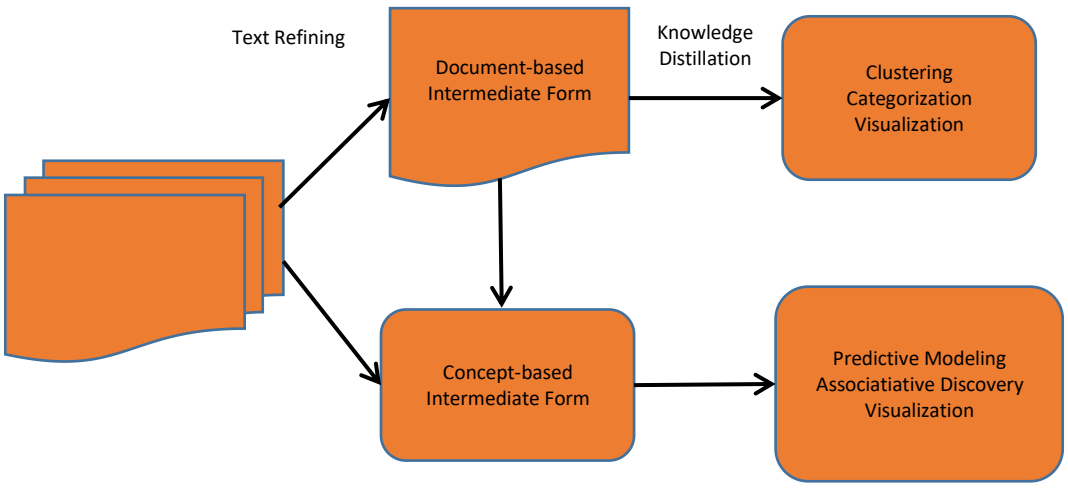
Knowledge Discovery in Text(KDT)[18] yang diusulkan oleh Feldman et al. Merupakan awal dari sistem Teks Mining dengan Arsitektur umum file seperti tampak pada gambar berikut :



Gambar 4 Sistem Arsitektur Knowledge Discovery in Text(KDT)

Pada Sistem ini input diambil dari dua variabel yang berbeda yaitu kumpulan dokumen berlabel kata kunci, dan kata kunci hierarki directed acyclic graph (DAG). Di sini semua istilah diidentifikasikan dengan nama-nama yang unik. Hierarki kata kunci diberi hierarki hubungan antara konsep utama yang terlibat dalam domain aplikasi. Modul operasi pencarian digunakan sebagai alat pengumpulan data yang memiliki kata kunci dengan tujuan mencari informasi baru, serta untuk menemukan pola baru yang dibutuhkan,

Pada Teks Mining Representatif yang dinyatakan olehTan[65] bahwa kerangka umum teks mining terdiri atas dua komponen yaitu : pemurnian teks dengan merubah bentuk dokumen teks bebas menjadi sebuah bentuk pilihan model yang ada, dan penyaringan informasi (*knowledge distillation*) yang menyimpulkan pola atau informasi dari bentuk model yang ada. Gambar 5 menunjukan sebuah Text Mining Framework



Gambar 5 menunjukan sebuah Text Mining Framework

Distilasi Informasi dari sebuah konsep berbasis pemilihan keputusan akan membentuk pola dan hubungan antar objek atau konsep. Mothe dkk.[47] menambahkan tumpukan data warehouse berdasarkan pola yang umum. Dimana seleksei Informasi memiliki tugas guna menyaring informasi yang sesuai dengan domain tujuan, yang harus diinformasikan. Format informasi harus disesuaikan, sebagai akibat dari banyaknya data yang tidak terstruktur. Dan pada bagian lain untuk menghindari adanya dampak dari imputasi nilai yang hilang, adanya nois dan dimensi, data yang tidak sama, serta untuk menangani pemrosesan dokumen yang formatnya tidak sesuai, maka Gracia, dkk.[59] menyarankan untuk menggunakan, ekstensi, dan popularitas pra-pemrosesan data berbasis algoritma.

Sementara itu Gürbüz, dkk.[21] menganjurkan agar mengecilkan data set dengan regresi analisis dan anomali deteksi analisis, namun teknik yang diusulkannya ini akan sulit jika diterapkan pada penanganan big data. Jamdar, dkk.[53] menganjurkan pembagian data untuk mempartisi data ke dalam partisi vertikal dan horizontal. Teknik pembagian ini memfasilitasi utilitas data yang baik dan hubungan antar file. Untuk pengelompokan dalam penggalian data Verma, dkk.[52] melakukan analisis terhadap enam algoritma pengelompokan seperti pengelompokan optik dan algoritma Expectation Maximization (EM), K-means, pengelompokan hierarki, pengelompokan berbasis kepadatan, Algoritma EM dan K-means efisien dalam menangani set data besar, pengelompokan DBScan . Lebih lanjut, ia menganalisis bahwa hierarkis algoritma pengelompokan sensitif terhadap nois.

Ada banyak model lain yang diusulkan untuk bidang aplikasi tertentu seperti Shehata dkk.[60], mengusulkan model mining baru berbasis konsep untuk pengelompokan teks. Model yang data inputnya adalah teks yang belum diproses, konsep ini meliputi analisis istilah berbasis konsep dan ukuran kesamaan berbasis konsep. Selanjutnya konsep ini mereka perluas [61] dengan menambahkan model analisis konsep berbasis kalimat, analisis konsep berbasis dokumen, analisis konsep berbasis korpus, dan berbasis konsep ukuran kesamaan, Pada konsep ini Model bisa secara efisien menemukan konsep pencocokan yang signifikan antara dokumen yang sesuai dengan semantik kalimat mereka.

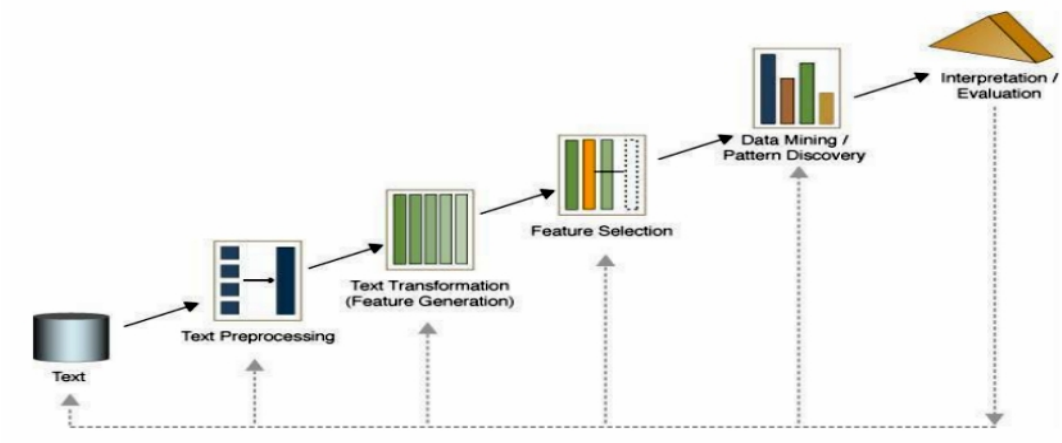
3.1. Pola Dan Algoritma Teks Mining

Sederhananya data mining merupakan proses mengekstraksi dan mengidentifikasi informasi atau *knowledge* dari tumpukan data yang banyak dan bersebaran pada sumber yang berbeda. Dalam prosesnya data mining menggunakan teknik statistik, matematika, dan

kecerdasan buatan. Istilah lain yang mengacu pada data mining adalah knowledge e xtraction, pattern analysis, information harvesting, dan data archaeology. Hasil data mining dapat digunakan untuk melakukan prediksi terhadap suatu masalah, menemukan informasi baru, menemukan pola yang belum diketahui, dan membantu dalam pengambilan keputusan.Teks mining sudah banyak memunculkan penelitian yang ahirnya membentuk bidang penerapan dan pola aplikasinya. Sesuai bidang penerapannya teks mining bisa dibagi menjadi "diklasifikasikan sebagai pengelompokan teks, pengelompokan teks, ekstraksi aturan asosiasi dan analisis tren".

3.2. Sistem Klasifikasi Teks

Didalam teks mining, sistem klasifikasi teks merupakan proses pengkajian dan evaluasi yang terpantau. Sistem klasifikasi teks merupakan sebuah proses penentuan untuk pemisahan dan pengelompokan teks secara langsung atau otomatis, sesuai dengan klasifikasi konten teks yang di bawah yang diberikansistem. Berikut ini adalah gambaran proses perjalanan teks mining [14]



Gambar 6. Teks Mining Proses

Penelitian pada bidang sistem klasifikasi teks ini sebenarnya sudah dimulai sekitar tahun 1960, khususnya pada indeks literatur ilmiah. Namun seiring dengan laju pertumbuhan jumlah data yang sangat signifikan serta meningkatnya permintaan untuk melakukan pemrosesan teks, ahirnya pada tahun 1990 sistem klasifikasi teks ini mulai dikembangkan secara total. Dalam beberapa tahun terakhir, sistem klasifikasi teks sudah digunakan secara otomatis dan semi otomatis pada banyak bidang, mulai dari deteksi genre, indeksasi hingga memfilter spam, pengertian kata yang memiliki banyak arti atau makna (disambiguasi)[24]-[15], pembuatan metadata, klasifikasi hirarki halaman-halaman web. Cara untuk mengenali disambiguasi yang biasanya dikaitkan dengan topik berbeda yang diusulkan oleh Gale dkk.[24] telah berhasil implementasikan pada identifikasi penulis dan pengambilan informasi, dengan menerapkan model keputusan bayesian kelas yang dimilikinya.

Analisis dari sebuah pandangan orang atau masyarakat adalah tergolong pada lingkup riset dibidang teks yang penting untuk dapat menghitung orientasi kata-kata yang mirip di dalam frasa berdasarkan prioritas bobot pandangan dan mengutamakan konsep untuk menghitung orientasi frasa kata tengah sesuai penggabungan kalimat dalam frasanya [13]. Lillian Lee dkk.[56] mengimplementasikan pola fitur bag-of standar serta tiga metode machine learning yaitu Naive Bayes, klasifikasi entropi maksimum, dan mendukung mesin vektor untuk pengelompokan sentimen, dengan film sebagai data ulasan.

Ada juga aplikasi yang digunakan sebagai filter email untuk membuang email "sampah (junk)"[1], mengidentifikasi teks bahasa untuk bahasa yang tidak dikenal[7], perkumpulan paten menjadi kategori-kategori memudahkan mereka dalam pencarian[43].

3.3. Clustering

Hipotesis cluster merupakan dasar pengelompokan teks, data-data yang masih berhubungan akan sangat mirip antara yang satu dengan yang lainnya, dibandingkan dengan data-data yang tidak berhubungan. Pengambilan dokumen, pencarian data, segmentasi citra, bidang analisis dan klasifikasi pola, merupakan kontribusi fungsi Pengelompokan (clustering). Ini banyak digunakan untuk meningkatkan kualitas presisi dan sistem penilaian ulang kualitas informasi atau output yang dihasilkan.[57]-[54].

Pengelompokan teks merupakan alat yang canggih dan paling cocok digunakan dalam melakukan analisis yang berhubungan dengan teks, dengan Pengelompokan teks kita bisa mendapatkan materi informasi yang kita butuhkan dalam dokumen dalam skala besar. Clifton,[9] dalam tulisannya menjelaskan metode analisis. Pertama ekstrak nama entitas dari semua dokumen yang ada, kemudian cari kumpulan item yang paling banyak atau paling sering muncul: kelompok entitas biasanya muncul bersama, perform selanjutnya nama kelompok entitas yang dikelompokkan berdasarkan frekuensi entitas yang paling sering dan paling banyak muncul menggunakan metode berbasis hypergraph [16]. Setiap cluster ditampilkan kembali sebagai sekumpulan entitas yang telah diberi nama dan sesuai dengan topik yang sedang berlangsung di dalam korpus (kumpulan teks yang digunakan untuk memperkuat analisis).

Untuk mendapatkan analisis informasi sebagai dampak dari sebuah informasi online yang sedang menjadi pembicaraan hangat pada media informasi Montes-y-Gómez dkk.[48] mengusulkan metode teks mining. Gejolak umum yang dapat kita perhatikan dalam berita adalah pengaruh dari pokok masalah atau topik berita yang sedang banyak dibicarakan ditengah masyarakat dan sempat menutupi berita lainnya namun dalam jangka pendek, atau biasa disebut dengan "*ephemeral association*" yaitu seuatu kejadian yang akan diingat orang namun hanya sementara. Mereka mengusulkan sebuah teknik dengan asosiasi yang dapat diamati dideteksi dengan metode statistik sederhana.

3.4. Association Rule Extraction

Agrawal [2] mengimplementasikan penggunaan metode Association Rule Extraction yang telah menjadi sebuah diskusi hangat didalam teks mining. Hal ini untuk melihat korelasi asosiasi antara atribut dan format kata yang berbeda dari sekumpulan teks yang ada.

Deskripsi formal dari aturan ekstraksi diberikan dalam asosiasi [49] seperti di bawah ini:
Diberikan kumpulan dokumen yang diindeks D = $\{d_1, d_2,\ldots,d_n\}$ dan satu set item A= $\{w_1, w_2,\ldots, w_n\}$, didalamnya terdapat frase atau konsep, istilah, dan kata kunci.
Biarkan W_i menjadi satu set item. Dokumen d_i dikatakan berisi W_i jika dan hanya jika $W_i \subseteq d_i$. Kaidah asosiasi merupakan implikasi dari bentuk $W_i \Longrightarrow W_j$, dimana $W_i \subset A$, $W_j \subset A$ dan $W_i \cap W_j = \phi$. Dua parameter penting yang menjadi standar yang mendasari aturan asosiasi yaitu dukungan dan kepercayaan. Aturan ekstraksi asosiasi dibagi menjadi dua langkah. Langkah pertama menghasilkan semua itemsets, disini dukungannya lebih besar dari pengguna dukungan minimum yang telah ditentukan dan disebut dengan "minupp". Set ini dikenal dengan istilah "frequent itemsets". Langkah kedua yaitu menggunakan frekuensi itemsets yang teridentifikasi untuk menghasilkan aturan yang memuaskan pengguna yang ditentukan keyakinan minimum yang disebut dengan "minconf". Dalam sistem teks mining terkadang dapat menghasilkan banyak sekali aturan asosias, akan tetapi hanya sedikit sekali dari aturan-aturan asosiasi tersebut disukai oleh end user (pengguna informasi). Maka sangatlah penting penilaian serta pemilihan aturan asosiasi ini pada sistem teks mining praktis. Referensi [3] memprediksi akan terjadi perkembangan pada aturan-aturan pencarian teks sebelumnya menggunakan ukuran jarak

semantik berdasarkan WordNet [19]. Jika semantik (makna kata dan kalimat) jaraknya pendek, maka aturan itu mungkin tidak berguna.

3.5. Tren Analisis

Apabila dimensi waktu dijadikan sebagai tolak ukur dari data teks, maka itu bisa menggambarkan perubahan aturan topik teks atau akan terjadi prediksi tren perkembangan objek perangkat [70]. Sekarang penelitian tentang tren analisis terutama diarahkan pada berita terbaru atau terkini seperti ; laporan keuangan, literatur ilmiah, informasi medis, laporan bisnis dan data teks penjadualan lainnya [27].

Mei, Qiaozhu, and C. X. Zhai [3] mengusulkan pendekatan probabilistik umum guna menemukan serta menyingkat pola tema dalam aliran teks dengan menemukan tema laten dari teks, membangun grafik evolusi tema, dan menganalisis kehidupan siklus tema. Untuk menemukan tema grafik evolusi, pertama-tama metode mereka akan menghasilkan kelompok kata (yaitu, tema) untuk setiap periode waktu dan kemudian menggunakan pengukuran divergensi (penyebaran) *Kullback-Leibler* untuk menemukan tema yang saling berhubungan (koheren) dari waktu ke waktu.

Grafik evolusi dapat mengungkapkan bagaimana tema berubah seiring waktu dan bagaimana satu tema dalam satu periode waktu memberi pengaruh tema lainnya pada periode tema selanjutnya. Mereka juga mengusulkan metode berdasarkan model Markov tersembunyi untuk menganalisis siklus hidup masing-masing tema. Pada metode ini pertama-tama akan menemukan tema-tema yang menarik secara global, kemudian melakukan perhitungan kekuatan masing masing tema pada setiap periode waktu. Hal ini memungkinkan Mei dkk. untuk membandingkan kekuatan relatif dari berbagai tema dari waktu ke waktu, bukan sekedar melihat tren variasi kekuatan tema.

Perubahan dari semua jenis paten selama bertahun-tahun akhirnya ditemukan melalui analisis paten databae yang relevan yang dilakukan oleh Brian Lent dkk.[39]. Victor Lavrenko dkk.[40] memprediksi tren harga saham berdasarkan berita dari perusahaan yang dikutip dan data historis harga saham. Sementara itu Montes-y-Gómez dkk. [51] menyajikan metode untuk menganalisis tren berita untuk menemukan berita terhangat di masyarakat saat ini dan kecenderungannya yang berubah. Saat ini, pekerjaan di bidang ini sebagian besar mengadopsi metode berdasarkan statistik [18][50]. Feldman dkk. [18] menggunakan kata kunci distribusi untuk memberi label pada dokumen, dan menghitung jarak antara distribusi kata kunci untuk koleksi dari berbagai titik waktu agar bisa menemukan tren perubahan topik teks.

4. KESIMPULAN

Telah penulis paparkan pengantar singkat untuk teks mining tersebut, Kemudian beberapa model umum yang telah dijelaskan untuk mengetahui perkembangan teks mining dalam keseluruhan perspektif. Dan pada bagian akhir penulis mengklasifikasikan pekerjaan teks mining sebagai kategorisasi teks, pengelompokan teks, aturan ekstraksi asosiasi dan tren analisis sesuai aplikasi. Perkembangan terbaru teks miningadalah masuk pada area kecerdasan buatan, dan dengan peningkatan teknologi teks mining berkelanjutan , perkembangan implementasi aplikasinyapun tumbuh.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Redaksi Jurnal JUPITER Politeknik Negeri Sriwijaya yang telah bersedia mempublish dan memberi dukungan terhadap makalah ini

DAFTAR PUSTAKA

[1] Androutsopoulos, Ion, et al. "An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages." Proceedings of Annual International Acm Sigir Conference on Research & Development in Information Retrieval(2000):160--167.

[2] Agrawal, Rakesh, T. Imieliński, and A. Swami. "Mining Association Rules Between Sets Of Items In Large Databases." SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data1993:207--216.

[3] Basu, Sugato, et al. "Using lexical knowledge to evaluate the novelty of rules mined from text." Proceedings of the NAACL workshop and other Lexical Resources:Applications, Extensions and Customizations. 2001.

[4] Berry, M. W., & Kogan, J. (2010). Text Mining Application and Theory. United Kingdom: WILEY.

[5] Beil, Florian, Martin Ester, and Xiaowei Xu. "Frequent term-based text clustering." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

[6] Berendt, Bettina, Andreas Hotho, and Gerd Stumme. "Towards semantic web mining." The Semantic Web—ISWC 2002. Springer Berlin Heidelberg, 2002. 264-278.

[7] Cavnar, William B., and J. M. Trenkle. "N-Gram-Based Text Categorization." Proceedings of Int'l Symposium on Document Analysis & Information Retrieval Las Vegas Nv(2001):161--175.

[8] Cooley, Robert, Bamshad Mobasher, and Jaideep Srivastava. "Web mining: Information and pattern discovery on the world wide web." Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on. IEEE, 1997.

[9] Clifton, Chris, and R. Cooley. TopCat: Data Mining for Topic Identification in a Text Corpus. Principles of Data Mining and Knowledge DiscoverySpringer Berlin Heidelberg, 1999:174-183.

[10] Dave, Kushal, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." Proceedings of the 12th international conference on World Wide Web. ACM, 2003.

[11] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, John W. Tukey "Scatter/Gather: a cluster-based approach to browsing large document collections." Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrievalACM, 1992:318--329.

[12] de Bruijn, Lambertus, and Joel Martin. "Literature mining in molecular biology." Proceedings of the EFMI Workshop on Natural Language Processing in Biomedical Applications. 2002.

[13] Dun LI, Fu-Yuan CAO, Yuan-Da CAO, Yue-Liang WAN. "Text Sentiment Classification Based on Phrase Patterns." Computer Science. 35.4(2008):132-134. DOI:10.3969/j.issn.1002-137X. 2008.04.037.

[14] Even, Yair and Zohar. Introduction to Text Mining.National Center for Supercomputing Applications Universty of Illinois.2002

[15] Escudero, Gerard, L. Marquez, and G. Rigau. "Boosting Applied to Word Sense Disambiguation." In Proceedings Of The 12th European Conference On Machine Learning 2000:129--141.

[16] E.-H. S. Han, G. Karypis, and V. Kumar, "Clustering Based on Association Rule Hypergraphs," Proc. SIGMOD'97 Workshop Research Issues in Data Mining and Knowledge Discovery, 1997.

[17] Feldman, Ronen, and I. Dagan. "Knowledge Discovery in Textual Databases (KDT)." In roceedings of the First International Conference on Knowledge Discovery and Data Mining DD-95(1995):112--117.

[18] Feldman, Ronen, I. Dagan, and H. Hirsh. "Mining Text Using Keyword Distributions. " Journal of Intelligent Information Systems10.3(1998):281-300.
[19] Fellbaum, C., and G. Miller. WordNet:An Electronic Lexical Database. MIT Press, 1998.
[20] Feldman, R., & Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. New York: Cambridge University Press.
[21] F. Gürbüz*, et al.*, "Data mining and preprocessing application on component reports of an airline company in Turkey," *Expert Systems with Applications,* vol. 38, pp. 6618-6626, 2011.
[22] Grace, L. K., V. Maheswari, and Dhinaharan Nagamalai. "Analysis of web logs and web user in web mining." arXiv preprint arXiv:1101.5668 (2011).
[23] Ghanem, M., Chortaras, A., Guo, Y., Rowe, A., & Ratcliffe, J. (2005). "A Grid Infrastructure For Mixed Bioinformatics Data And Text Mining." Computer Systems and Applications, 2005. The 3rd ACS/IEEE International Conference on(Vol.29, pp.41-I)
[24] Gale, W. A., Church, K. W., and Yarowsky, D. (1993). "A Method for Disambiguating Word Senses in a Large Corpus." Computers and the Humanities 26(5): 415–439
[25] Hu, Qinghua, et al. "A novel weighting formula and feature selection for text classification based on rough set theory." Natural Language Processing and Knowledge Engineering, 2003. proceedings. 2003 International Conference on IEEE, 2003:638-645.
[26] Hotho, Andreas, Steffen Staab, and Gerd Stumme. "Ontologies improve text document clustering." Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003.
[27] Hemalatha, I., Varma P. Saradhi, & Govardhan A. (2012). Preprocessing The Informal Text for Efficient Sentiment Analysis. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 1, 58-61
[28] He, X., Zhang, S. & Liu, Y., 2015. An Adaptive Spectral Clustering Algorithm Based on the Importance of Shared Nearest Neighbors. algorithms, 8, pp.177–189.
[29] http://www-01.ibm.com/common/ssi/cgibin/ssialias?infotype=an&subtype ca&htmlfid =897/ENUS298-061&appname=isource&language=enus#3pb
[30] http://www.sas.com/en_us/software/analytics/text-miner.html
[31] http://www.textanalysis.com/Products/VisualText/visualtext.html
[32] Hearst, Marti A. "Untangling text data mining." University of Maryland1999:3-10.
[33] Jiang, Chuntao, et al. "Text Classification using Graph Mining-based Feature Extraction." Knowledge-Based Systems23.4(2010):302–308.
[34] J. W. Reed*, et al.*, "A multi-agent system for distributed cluster analysis," in *Proceedings of Third International Workshop on Software Engineering for Large-Scale Multi-Agent Systems (SELMAS'04) Workshop in conjunction with the 26th International Conference on Software Engineering Edinburgh, Scotland, UK: IEE*, 2004, pp. 152-5.
[35] Kosala, Raymond, and Hendrik Blockeel. "Web mining research: A survey." ACM Sigkdd Explorations Newsletter 2.1 (2000): 1-15.
[36] Karanikas, Haralampos, C. Tjortjis, and B. Theodoulidis. "An Approach to Text Mining using Information Extraction." Proc. Workshop Knowledge Management Theory Applications (KMTA 00(2000).
[37] Luhn, H. P. "A Business Intelligence System." Ibm Journal of Research & Development2.4 (1958):314-319.
[38] Larsen, Bjornar, and C. Aone. "Fast and effective text mining using linear-time document clustering." Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data miningACM, 1999:16-22.

[39] Lent, Brian, Rakesh Agrawal, and Ramakrishnan Srikant. "Discovering Trends in Text Databases." KDD. Vol. 97. 1997.
[40] Lavrenko, Victor, et al. "Mining of Concurrent Text and Time Series." Proceedings of Acm Sigkdd Intl Conference on Knowledge Discovery & Data Mining Workshop on Text Mining(2000):37--44.
[41] Liu, B. (2010). Handbook of Natural Lenguage Processing 2nd Edition. Boca Raton: CRC Press.
[42] Luo, Congnan, Y. Li, and S. M. Chung. "Text document clustering based on neighbors.." Data & Knowledge Engineering68.11(2009):1271-1288.
[43] Larkey, Leah S. "A Patent Search and Classification System." Digital Libraries the Fourth Acm Conference on Digital Libraries(1999):79--87.
[44] Masoud Makrehchi and Mohamed S. Kamel. "Text Classification Using Small Number of Features.." Lecture Notes in Computer Science(2005):580-589.
[45] Mittermayer, Marc-André. "Forecasting intraday stock price trends with text mining techniques." System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on. IEEE, 2004.
[46] Maron, M. E., and J. L. Kuhns. "On Relevance, Probabilistic Indexing and Information Retrieval.." Journal of the Acm7.3(1960):216-244
[47] Mothe J., Chrisment C., Dkaki T., Dousset B., Egret D., (2001) "Information mining: use of the document dimensions to analyse interactively a document set", European Colloquium on IR Research: ECIR, 66-77.
[48] M. Montes-y-Gómez, A. Gelbukh, and A. López-López. "Discovering Ephemeral Associations among News Topics." 17th international joint conference on artificial intelligence ijcai-01, workshop on adaptive text mining 2001.
[49] Mahgoub, Hany, et al. "A Text Mining Technique Using Association Rules Extraction." International Journal of Computational Intelligence1(2008):21.
[50] Mei, Qiaozhu, and C. X. Zhai. "Discovering evolutionary theme patterns from text: an exploration of temporal text mining." Proceedings of Kdd '(2005):198-207.
[51] Montes-y-Gómez, López –López and Gelbukh, "Text Mining as a Social Thermometer", Proc. Of the Workshop on Text Mining: Foundations, Techniques and Applications, IJCAI-99, Stockholm, 1999.
[52] M. Verma*, et al.*, "A comparative study of various clustering algorithms in data mining," *International Journal of Engineering Research and Applications (IJERA),* vol. 2, pp. 1379-1384, 2012.
[53] N. Jamdar and V. Babane, "Survey on Privacy-Preservation in Data Mining Using Slicing Strategy," *International Journal of Science and Research (IJSR),* vol. 2, pp. 306-309, 2013.
[54] Oren Zamir, Oren Etzioni, Omid Madani, Richard M. Karp. "Fast and Intuitive clustering of Web documents." Proceedings of International Conference on Knowledge Discovery & Data Mining (1997):287--290.
[55] Patidar, A.K., Agrawal, J. & Mishra, N., 2012. Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbor Clustering Approach. International journal of Computer application, 40(16), pp.1–5.
[56] Pang, Bo, L. Lee, and S. Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques." Proceedings of Emnlp(2002):79--86.
[57] Rijsbergen, Van. C.J. "Information Retrieval." 14th International Symposium on Methodologies for Intelligent Systems. Volume 2871., Maebashi City, Japan, LNCS, Springer-Verlag 12.2-3(1989):95.
[58] R. Burke, M.P. O'Mahony, N.J. Hurley, Robust collaborative recommendation, Recommender Systems Handbook, Springer, 2015, pp. 961–995

[59] S. García*, et al.*, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," *Knowledge-Based Systems,* 2015.
[60] Shehata, S., F. Karray, and M. Kamel. "Enhancing Text Clustering Using Concept-based Mining Model." IEEE 13th International Conference on Data MiningIEEE Computer Society, 2006:1043-1048.
[61] Shehata, Shady, F. Karray, and M. S. Kamel. "An Efficient ConceptBased Mining Model for Enhancing Text Clustering." IEEE Transactions on Knowledge & Data Engineering 22.10(2010):1360-1371.
[62] S. Grimes. "Unstructured data and the 80 percent rule." Carabridge Bridgepoints, 2008.
[63] Salton, G., A. Wong, and C. S. Yang. "A vector space model for automatic indexing." In Communications of the ACM 18(11), 1975: 613-620.
[64] Steinheiser, R., and C. Clifton. "Data Mining on Text." 2012 IEEE 36th Annual Computer Software and Applications ConferenceIEEE Computer Society, 1998:630.
[65] Tan, Ah Hwee, et al. "Text Mining: The state of the art and the challenges." Proceedings of the Pakdd Workshop on Knowledge Disocovery from Advanced Databases(2000):65--70.
[66] Tan, Songbo, et al. "Using dragpushing to refine centroid text classifiers." Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005.
[67] Tanabe, L., et al. "MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling." Biotechniques 27.6 (1999): 1210-4.
[68] Wu, F. et al., 2015. A Nearest Neighbor Searches (NNS) Algorithm for Fast Registration of 3D Point Clouds based on GPGPU. In International Conferences on Intelligent Research and Mechatronics Engineering (ISRME). pp. 2153–2158.
[69] Witten, Ian H., et al. "Text mining in a digital library." International Journal on Digital Libraries 4.1 (2004): 56-59.
[70] Zhi-Qun Chen, and Guo-Xuan Zhang. "A Survey of Text Mining." Journal of Pattern recognition and artificial intelligence 18.1(2005):65-74. DOI:10.3969/j.issn. 1003-6059.2005.01.012.
[71] Zhi-Qun Chen. " A Survey of Trend Mining for texts." Journal of Information Science 2(2010):316-320.
[72] Zahrotun, L., 2016. Comparison Jaccard similarity , Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method. , 5(1), pp.11–18.