Topic Discovery pada Dokumen Abstrak Jurnal Penelitian di *Science Direct* Menggunakan *Association Rule*

Mochammad Farros Fatchur Roji dan Irhamah
Departemen Statistika, Fakultas Matematika, Komputasi, dan Sains Data,
Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
e-mail: irhamah@statistika.its.ac.id

Abstrak*— Jurnal memiliki peranan penting dalam meningkatkan pemahaman mengenai ilmu berdasarkan review dari ilmuwan. Karakteristik jurnal seperti update berkaitan dengan teori dibandingkan buku, pembahasan yang lebih ringkas, sebagai referensi alternative, aplikasi dan implementasi dunia nyata. Jurnal yang telah dibuat dalam bentuk digitalisasi memiliki istilah lain yaitu file atau soft copy dengan memanfaatkan teknologi informasi dan komunikasi, yang saat ini menjadi salah satu koleksi perpustakaan digital. Data yang di gunakan berasal dari ScienceDirect. ScienceDirect adalah database yang berisi kumpulan dokumen full-text yang berkualitas yang telah diperiksa oleh peer-review Elsevier. Dokumen abstrak dari sciencedirect tersebut nantinya akan dilakukan pre processing terlebih dahulu. Kemudian di lanjutkan dengan association rule dan pearson correlation setelahnya. Pada association rule term kata jika menggunakan min support 2 % maka di dapatkan frequent itemset sebanyak 72, closed frequent itemset sebanyak 55, dan remove subset sebanyak 41 itemset. Kemudian saat di lakukan analisis korelasi pada itemset remove subset. Di dapatkan bayesian,model yaitu itemset yang paling banyak memiliki hubungan. Kemudian pada topic community dengan cfinder terbagi menjadi dua komunitas dan terdapat irisan sebanyak 6 itemset.

Kata Kunci*— Association Rule, E-Jurnal, Pearson Correlation, Pre processing, ScienceDirect

I. PENDAHULUAN

Jurnal adalah publikasi yang membahas berbagai macam ilmu pendidikan serta penelitian yang memiliki interval jangka waktu terbit berkesinambungan. Sedangkan elektronik jurnal adalah publikasi jurnal yang sudah dikemas dalam bentuk digitalisasi. Jurnal memiliki peranan penting dalam meningkatkan pemahaman mengenai ilmu berdasarkan review dari ilmuwan. Karakteristik jurnal seperti update berkaitan dengan teori dibandingkan buku, pembahasan yang lebih ringkas, sebagai referensi alternative, aplikasi dan implementasi dunia nyata. Berbicara e-journals dalam bukunya Putu Laxman Pendit dijelaskan, Tahun 1990-an e-journals masih dalam tahap eksperimental dan masih sesuatu mimpi, tapi sekarang terbukti semua itu bukan lagi mimpi telah menjadi kenyataan diluar negeri seperti EBSCO, OCLC, PROQUEST, I-GROUPS adalah perusahaan-perusahaan yang mengelola informasi jurnal menjadi lebih menarik dan mudah diakses.

E-Jurnal merupakan bentuk digitalisasi istilah lain file atau soft copy dengan memanfaatkan teknologi informasi dan komunikasi, yang saat ini menjadi salah satu koleksi perpustakaan digital. Pengadaan koleksi tersebut di sebuah perpustakaan tujuan akhirnya adalah untuk kepuasan pengguna khususnya anggota. Yaitu dengan memberikan layanan yang tidak terbatas oleh ruang dan waktu dengan media internet. Bentuk elektronik jurnal di akses dengan media internet untuk di kampus maupun di luar kampus, dan mengantisipasi koneksi internet down, telah disediakan akses offline di kampus [2].

Penelitian sebelumnya mengenai penelitian dokumen abstak pernah di lakukan oleh Jose L. Hurtado, Ankur Agarwal dan Xingquan Zhu yang berjudul *Topic discovery and future trend forecasting for texts*. Dalam jurnal tersebut di lakukan dengan mengumpulkan topik dari sekumpulan dokumen, seperti jurnal penelitian, hak cipta, dan laporan teknis, yang sangat membantu untuk meringkas data dokumen dalam skala besar juga membantu meramalkan tren topik di masa depan. Bermanfaat juga bagi banyak aplikasi, seperti memodelkan evolusi arah penelitian dan meramalkan tren masa depan industri teknologi informasi. Penelitian tersebut menggunakan analisis asosiasi dan peramalan ensemble untuk secara otomatis menemukan topik dari serangkaian dokumen teks dan memperkirakan tren topik yang berkembang dalam waktu dekat. Pengumpulan publikasi tersebut berasal dari area penelitian, *data mining*, dan *machine learning* sebagai domain data. Jadi pertama-tama di lakukan mengidentifikasi serangkaian topik untuk analisis asosiasi, diikuti oleh analisis korelasi untuk membantu menemukan korelasi antar topik, dan mengidentifikasi jaringan topik dan komunitas.

Dokumen abstrak bersumber dari *ScienceDirect* yang nantinya akan di Analisa menggunakan *association rule*. Analisis asosiasi atau *association rule mining* adalah teknik data mining untuk menemukan aturan asosiatif antara suatu kombinasi item. Analisis asosiasi dikenal juga sebagai salah satu teknik data mining yang menjadi dasar dari salah satu teknik data mining lainnya. Secara khusus, salah satu tahap analisis asosiasi yang menarik perhatian banyak peneliti untuk menghasilkan algoritma yang efisien, yaitu analisis pola frekuensi tinggi (frequent pattern mining) [5].

Detail dokumen abstrak pada *ScienceDirect* tersebut nantinya akan dilakukan *pre processing* terlebih dahulu, adapun tahapannya dari *text pre processing* adalah *tokenization* yaitu memecah berita menjadi kata per kata dan *case folding untuk* mengubah semua teks dengan huruf kecil serta menghilangkan tanda baca*, removing stops words and small words, part of speech tagger (POS)* yaitu pengkatogorian kata*, removing verbs, stemming* (menemukan kata dasar dari sebuah kata), dan *lemmatization* (menghasilkan kata dasar dengan memperhatikan kamus). Kemudian setelah dilakukan *text pre processing* maka di lanjutkan dengan *association rule mining* untuk menemukan aturan asosiatif antara suatu kombinasi item juga untuk mencari kandidat topik. Kemudian dari kandidat topik tersebut dilakukuan metode *closed frequent itemset* supaya topiknya terdiri dari jumlah maksimum kata kunci umum yang ada di semua dokumen dan analisis *pearson correlation* untuk mengetahui hubungan antara satu variabel dengan variabel yang lain secara linier. Kemudian dari korelasi

tersebut bisa terbentuk *topic network & community* yang kemudian di lanjut dengan regresi *trend.*

II. TINJAUAN PUSTAKA

A. Text Preprocessing

Data prepocessing merupakan tahapan-tahapan yang dilakukan sebelum mengolah data yang telah diperoleh agar bisa dilakukan analisis lebih lanjut. *Text preprocessing* bertujuan untuk mengubah data textual yang tidak berstruktur ke dalam data yang terstruktur dan disimpan dalam basis data. Adapun tahapan-tahapan praproses data adalah sebagai berikut.

1. *Tokenization* dan *case conversion*

   Tokenization adalah proses untuk membagi teks input menjadi unit-unit kecil yang disebut token [4]. Token atau biasa disebut juga term bisa berupa suatu kata, angka atau tanda baca. Pada penelitian ini tanda baca dihilangkan sehingga tidak dianggap sebagai token. Sedangkan *Case folding* dilakukan untuk menghilangkan karakter selain huruf dan mengubah seluruh huruf menjadi *lowercase.*
2. *Removing stops words and small words.*

   Stopword merupakan kata yang sering muncul dalam dokumen seperti "between", "and", "this", "on", "an", "a", "the", dll. kata-kata yang masuk dalam stopword seringkali dianggap tidak memiliki makna, sehingga kata yang tercantum dalam daftar ini dibuang dan tidak ikut diproses pada tahap selanjutnya.
3. *Part of speech Tagger (POS)*

   Proses pengelompokan kata ke dalam *part of speech* dan pelabelan yang sesuai dikenal sebagai Part of Speech Tagging. *part of speech* juga dikenal sebagai kelas kata atau kategori leksikal. Algoritma yang digunakan dalam melakukan labelling kata adalah menggunakan Natural Language Toolkit (NTLK).
4. *Removing verbs.*

   Removing verbs adalah menghapus kata dengan label *verb. verb* membantu dalam menggambarkan *action* atau tindakan tetapi tidak menjelaskan topik itu sendiri dan karena itu *verb* / kata kerja dihapus.
5. *Stemming (reduce to word stem).*

   Untuk alasan tata Bahasa atau dalam Bahasa Inggris disebut *grammar*, dalam satu dokumen akan menggunakan berbagai bentuk kata, seperti *organize, organizes, dan organizing.* Selain itu ada kelompok kata yang merupakan turunan dari kata dasar yang mempunyai makna yang sama, seperti *democracy, democratic,* dan *democratization*. Dalam banyak situasi, pencarian kata dasar untuk kata-kata yang mempunyai makna yang sama akan sangat berguna [4].

   Stemming merupakan suatu proses untuk menemukan kata dasar dari sebuah kata. Dengan menghilangkan semua imbuhan (*affixes*) baik yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan *confixes* (kombinasi dari awalan dan akhiran) pada kata turunan. NLTK menyediakan beberapa interface *stemmers* terkenal, seperti *Porter stemmer*, *Lancaster Stemmer*, *Snowball Stemmer* dan lain-lain.
6. *Lemmatization*

   Lemmatization hampir sama seperti *stemming*, yang membedakan adalah pada *stemming* lebih banyak memotong akhir kata, dan sering juga membuang imbuhan tetapi pada *lemmatization* menghasilkan kata dasar dengan memperhatikan kamus. Sebagai contoh kata "studies" pada *stemming* menghasilkan output "studi" sedangkan pada *lemmatization* menghasilkan output "study".

B. Association Rule

Analisis asosiasi atau *association rule mining* adalah teknik data mining untuk menemukan aturan asosiatif antara suatu kombinasi item. Analisis asosiasi dikenal juga sebagai salah satu teknik data mining yang menjadi dasar dari salah satu teknik data mining lainnya. Secara khusus, salah satu tahap analisis asosiasi yang menarik perhatian banyak peneliti untuk menghasilkan algoritma yang efisien, yaitu analisis pola frekuensi tinggi (frequent pattern mining) [10].

Secara umum association *rule* mempunyai bentuk : LHS => RHS dimana LHS dan RHS tersebut adalah himpunan item; jika setiap item-item dalam LHS terdapat dalam transaksi maka item-item dalam RHS juga terdapat dalam transaksi.

Aturan asosiasi biasanya dinyatakan dalam bentuk sesuai persamaan (1):

$$\{A, B\} => \{C\}\ (support(A, B) = 10\%, confidence(A, B) = 50\%). \quad (1)$$

Support dari suatu *association rule* adalah presentasi kombinasi item tersebut dalam database, dimana jika mempunyai kata A dan kata B maka *support* adalah proporsi dari transaksi dalam *database* yang mengandung kata A dan kata B. Rumus untuk menghitung nilai *support* dari dua kata tersebut adalah sesuai persamaan (2) dan (3):

$$Support\ (A, B) = P(A \cap B), \quad (2)$$

$$Support(A, B) = \frac{\sum Transaksi\ mengandung\ A\ dan\ B}{\sum Transaksi}. \quad (3)$$

Confidence dari *association rule* adalah ukuran ketepatan suatu *rule*, yaitu presentasi dalam database yang mengandung kata A dan mengandung kata B. Dengan adanya *confidence* kita dapat mengukur kuatnya hubungan antar-item dalam association *rule*. Rumus untuk menghitung nilai *confidence* dari dua item tersebut adalah sesuai persamaan (2.4) dan (2.5)[6]:

$$Confidence(A, B) = P(B|A), \quad (4)$$

$$Confidence(A, B) = \frac{\sum Transaksi\ mengandung\ A\ dan\ B}{\sum Transaksi\ mengandung\ A} \quad (5)$$

1. Apriori

   Algoritma apriori adalah langkah untuk proses menemukan *frequent-itemset* dengan melakukan iterasi pada data. Dimana *itemset* adalah himpunan item-item yang berada di dalam himpunan yang diolah oleh sistem, sedangkan *frequent-itemset* menunjukan *itemset* yang memiliki frekuensi kemunculan lebih dari nilai minimum yang telah ditentukan (ϕ). Pada iterasi ke-k, semua *itemset* yang ditemukan yang memiliki k item disebut *k-itemset*. Setiap iterasi terdiri dari dua tahap yaitu pembangkitan kandidtat dan pembangkitan *rule*.

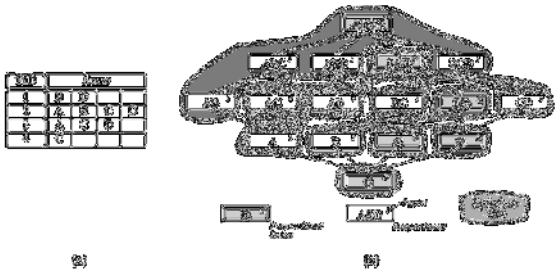
   Pada tahap pembangkitan kandidat (*candidat generation*) dimana himpunan semua *frequent*$(k-1)$-*itemset* F_{k-1} yang digunakan pada pass ke-$(k-1)$ digunakan untuk membangkitkan kandidat *itemset* C_k. Prosedur pembangkitan kandidat menjamin bahwa C_k adalah superset dari himpunan. Kemudian di-scan dalam tahap perhitungan *support (support counting)*. Pada akhir pass C_k diperiksa untuk menentukan kandidat mana yang muncul, menghasilkan F_k. Perhitungan *support* berakhir ketika C_k atau C_{k+1} kosong.

   Pada tahap membangkitkan *rule*, akan dibangkitkan lebih dahulu *candidate rule*. *Candidate rule* berisi semua kemungkinan *rule* yang memiliki *support* > *minimum support* karena inputan *candidate rule* adalah *frequent-itemset*. Kemudian candidate *rule* akan di-join dengan table F untuk menemukan *support antecedent*. *Confidence rule* dihitung dengan cara membandingkan *support rule* dengan *support antecendent rule*. Hanya *rule* yang mempunyai *confidence* > *minimum confidence* yang disimpan dalam *table rule* (table R)[7].

2. *Closed Frequent Keyword-sets as Topics*

Closed frequent itemset atau penutupan frekuensi set kata kunci sebagai topik untuk efisiensi dari *association rules* [9]. Pada penelitian ini, terdapat rangkaian topik yang sering muncul. Topik tersebut berasal dari set kata kunci yang sama. Dalam algoritme ini, kita dapat memperoleh set kata kunci topik yang sering tidak ditutup. Misalnya, ABCD dan ABC mungkin memiliki makna topik yang sama dalam kelompok tersebut. Dalam hal ini kita dapat menghapus ABC karena sudah memiliki informasi yang dimiliki oleh ABCD. Topik kita harus terdiri dari jumlah maksimum kata kunci umum yang ada di semua dokumen tersebut. Karenanya kita harus menutup beberapa set kata kunci sebagai topik. Misal adalah melalui perhitungan dari rangkaian kata kunci yang sering dalam *mode top-down* dalam satu pemindaian. Jadi kita tidak dapat menghilangkan set kata kunci yang tidak tertutup dalam prosedur 1 itu sendiri.

Untuk menghilangkan set kata kunci yang sering tidak ditutup maka di lakukan pengulangan *level-wise* dalam daftar set kata kunci yang sering. Dilakukan penyimpanan set kata kunci yang sering dalam *level-wise*, dengan jumlah kata kunci dalam set kata kunci yang mewakili tingkatnya. Untuk setiap set kata kunci dengan panjang i, maka dilakukan daftar set kata kunci dengan panjang (i + 1) dan jika set kata kunci *i-length* adalah *substring* dari (i + 1) panjang kata kunci-set dan dukungan sama untuk keduanya, maka dilakukan penghapusan set kata kunci i-*length*, karena tidak tertutup [10].



(Sumber : https://www.computer.org/csdl/trans/tk/2006/01/k0021.html)

Gambar 1 *Closed Frequent Keyword-sets as Topics*

C. *Pearson Correlation*

Analisis korelasi Pearson *(Correlate Bivariate)* digunakan untuk mengetahui hubungan antara satu variabel dengan variabel yang lain secara linier. Data yang digunakan berskala interval atau rasio. Nilai korelasi (*r*) adalah 0 sampai 1, semakin mendekati 1 hubungan yang terjadi semakin kuat. Sebaliknya, nilai semakin mendekati 0 maka hubungan yang terjadi semakin lemah [8]. Rumus ini disebut juga koefisien korelasi Pearson (*Pearson's product moment coefficient of correlation*).

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i{}^2 - \left(\sum_{i=1}^{n} X_i{}^2\right)^2}\sqrt{n\sum_{i=1}^{n} Y_i{}^2 - \left(\sum_{i=1}^{n} Y_i{}^2\right)^2}} \quad (6)$$

Dengan

$$t_{hitung} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (7)$$

$$t_{tabel} = t_{\left(\frac{\alpha}{2}, n-2\right)} \quad (8)$$

Pengujian menggunakan distribusi t sebagai uji statistiknya.

Prosedur pengujiannya adalah :

1. Menentukan formula hipotesis

   $H_0 : \rho = 0$ (tidak ada hubungan antara X dan Y)

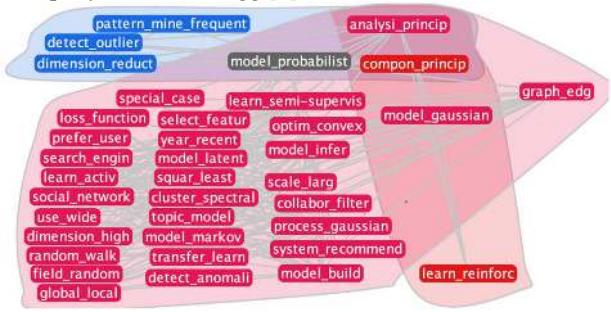
   $H_1 : \rho \neq 0$ (ada hubungan antara X dan Y)
2. Menentukan taraf nyata (α)
3. Menentukan kriteria penolakan

   Jika *P-value* yang dihasilkan lebih dari α maka gagal tolak H_o artinya tidak ada hubungan antara variabel X dan Y sebab nilai t hitung kurang dari nilai t tabel.

   Jika *P-value* yang dihasilkan kurang dari α maka tolak H_o artinya ada hubungan antara variabel X dan Y sebab nilai t hitung lebih dari nilai t tabel.
4. Membuat kesimpulan

   Menyimpulkan H_0 gagal tolak atau ditolak

Nilai koefisien korelasi ini digunakan untuk menentukan kumpulan topik-topik yang mempunyai korelasi tinggi atau saling berhubungan. Koefisien korelasi antar topik digunakan untuk *Network Analysis,* dengan masing-masing *node* menyatakan topik dan *edge* antar *node* menyatakan nilai koefisien korelasi antara dua topik yang mempunyai nilai diatas *threshold*. Gambar 2 merupakan contoh dari *Network Analysis* untuk menentukan pengelompokan topik-topik yang mempunyai korelasi tinggi[1].



Gambar 2 Topic Network and Community

D. Analisis Regresi

Analisis regresi bertujuan untuk mengestimasi atau menduga suatu hubungan antara variabel tahun, misalnya Y = f(x). Kedua, melakukan peramalan atau prediksi nilai variabel terikat (tidak bebas) atau dependent variable berdasarkan nilai variabel terkait (variabel independen/bebas).

Pendugaan parameter. Bila diberikan data contoh $\{(x_i, y_i); i=1,2,.....,n\}$. Dapat diperoleh dari rumus (9) dan (10)

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad (9)$$

$$a = \bar{y} - b\bar{x} \quad (10)$$

Keterangan:

b = nilai dugaan kuadrat terkecil bagi parameter

x_i = nilai data x ke-i.

y_i = nilai data y ke-i.

n = banyaknya data.

Jika ditaksir oleh a dan b, maka regresi linier berdasarkan sampel pada rumus (11).

$$\hat{y} = a + bx \quad (11)$$

Keterangan :

$\hat{y}$ = nilai yang diukur/dihitung pada variabel tidak bebas

x = nilai tertentu dari variabel bebas

a = intersep/perpotongan garis regresi dengan sumbu y.

b = koefisien regresi/kemiringan dari garis regresi/untuk mengukur kenaikan atau penurunan y untuk setiap perubahan satu-satuan x / untuk mengukur besarnya pengaruh x terhadap y kalau x naik satu unit[11].

III. METODOLOGI PENELITIAN

A. Sumber Data

Sumber data yang akan digunakan dalam penelitian ini adalah dokumen abstrak yang berasal dari jurnal-jurnal di *sciencedirect.com* dengan pengambilan keyword yang di cari adalah data mining karena topik tersebut memiliki hubungan dengan metode yang di bahas pada penelitian ini yaitu tentang text mining. Kemudian waktu terbit jurnal yang di gunakan adalah sebelas tahun (2008-2018).

B. Variabel Penelitian

Variabel yang di peroleh setelah dilakukan *crawling* dari website *ScienceDirect.com* adalah terdiri dari label, tipe, penulis, tahun terbit, judul, tema, volum, isu, jumlah halaman, abstrak, DOI, kata kunci, ISSN, dan URL *Langkah Analisis*

Langkah analisis digunakan untuk menggambarkan langkah-langkah penelitian yang akan dilakukan secara urut. Langkah analisis yang digunakan adalah sebagai berikut.

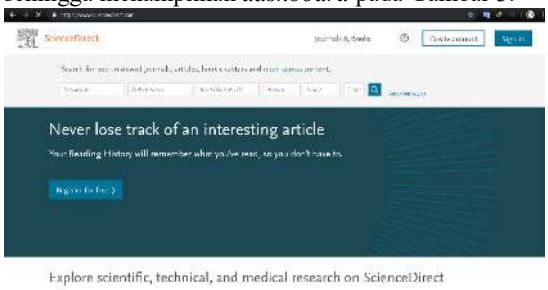
1. Menyiapkan data *keywords* yang di ambil dari abstrak *ScienceDirect.com*. Melakukan login akun *ScienceDirect* terlebih dahulu supaya mampu mengunduh banyak abstrak sekaligus dalam melakukan *web crawling.*
2. Mencari topik dengan *association rule* menggunakan 1-kata sebagai *1- item* dengan langkah-langkah sebagai berikut.
   a. Melakukakn *case folding, remove specific number, & remove punctuation*, yaitu memecah berita menjadi kata per kata dan mengubah semua teks dengan huruf kecil serta menghilangkan tanda baca serta angka spesifik.
   b. Menghapus kata yang mengandung *stopwords*,
   c. Melakukan *Lemmatization,* hampir sama seperti *stemming*, yang membedakan adalah pada *stemming* lebih banyak memotong akhir kata, dan sering juga membuang imbuhan tetapi pada *lemmatization* menghasilkan kata dasar dengan memperhatikan kamus.
   d. Melakukan visualisasi menggunakan wordcloud setelah di lakukan *preprocessing* serta menggunakan *wordcloud* n-gram yang terdiri atas 4 yaitu *wordcloud* 1-gram hingga 4-gram. Ditampilkan juga *bar-chart* untuk mengetahui frekuensi kata yang paling banyak.
   e. Melakukan *association rule* untuk menemukan aturan asosiatif antara suatu kombinasi item dengan tujuan mencari kandidat topik menggunakan algortima *apriori*.
   f. Melakukan *closed frequent itemset* untuk mendapatkan efisiensi dari kandidat topik yang di dapatkan dari *association rule* ditambah dengan *remove subset* untuk menghapus himpunan bagian dari itemset dan mengekstrasinya.
   g. Melakukan analisis korelasi pearson *(Correlate Bivariate)* untuk mengetahui hubungan antara satu variabel dengan variabel yang lain secara linier.
   h. Visualisasi hasil korelasi ke dalam bentuk *topic network* dan *topic community*
   i. Melakukan analisis regresi untuk mengetahui pola *trend* pertahunnya.
3. Mencari topik dengan *association rule* menggunakan 1-kata sebagai *1- item* dengan langkah-langkah:
   a. Melakukakan *case folding* untuk mengubah semua teks dengan huruf kecil.
   b. Melakukan *association rule* untuk menemukan aturan asosiatif antara suatu kombinasi item untuk mencari kandidat topik dengan algortima apriori.
   c. Ekstrasi keywords hasil dari *association rule.*
4. Membuat kesimpulan hasil analisis.

IV. ANALISIS DAN PEMBAHASAN

A. Tata Cara Pengambilan Data Keyword

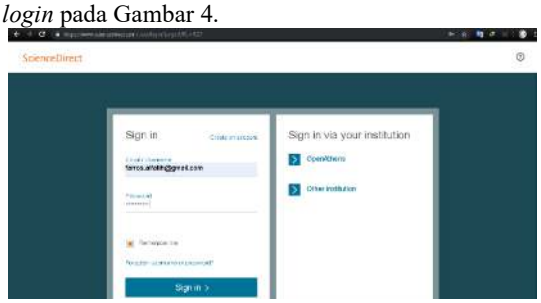
Data yang digunakan pada penelitian ini terdiri dari 2826 jurnal yang diambil dari jurnal-jurnal di *ScienceDirect repository*. Sebelum melakukan analisis, perlu dijelaskan terlebih dahulu bagaimana cara melakukan pengambilan data *keyword* karena banyak orang yang masih belum tahu. Berikut merupakan tata cara pengambilan abstrak di *ScienceDirect repository.*

1. Buka alamat website di https://ScienceDirect.org sehingga menampilkan *dashboard* pada Gambar 3.



Gambar 3. *Dashboard Website DataScience*

2. Klik pada tombol login terlebih dahulu di pojok kanan atas dengan pendaftarannya secara gratis sebab agar mampu mendownload banyak abstrak sekaligus harus login terlebih dahulu. Dengan tampilan *dashboard login* pada Gambar 4.



Gambar 4. *Dashboard Login.*

3. Ketikkan alamat email dan *password* di kolom login. Isi dengan benar kemudian di klik pada tombol *sign in* hingga keluar *dashboard* pada Gambar 5 yang menandakan bahwa kita telah berhasil *login*.

Gambar 5. *Dashboard Website* setelah *Login*

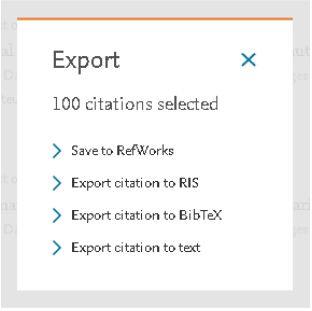
4. Kita isikan pada kotak *Journal/book title* yaitu topik jurnal yang ingin di cari. Pada penelitian kali ini yang menjadi topik adalah *Computational Statistics & Data Analysis*. Kemudian setelah topik jurnal sudah di inputkan kita klik tombol *search* warna biru di samping kanannya agar di proses oleh website ScienceDirect.com dengan tampilan selanjutnya bisa di lihat pada Gambar 6 berikut.

Gambar 6. *Dashboard* hasil pencarian

5. Centang kotak-kotak atas data yang kita pilih yaitu pada tahun dan juga jenis artikel yaitu *research articles* pada letak kiri dari tampilan website ScienceDirect.org serta kita klik kotak dialog kecil di samping tulisan *Download selected articles* untuk mendownload semua jurnal yang terletak pada halaman tersebut kemudian kita klik pada tombol *export* dengan tampilan pada Gambar 7.

Gambar 7. *Dashboard checklist* pada jurnal.

6. klik pada tombol *Export* sehingga muncul kotak dialog seperti Gambar 8.



Gambar 8. *Dashboard Export* Jurnal Penelitian.

7. Setelah itu kita pilih *Export citation to RIS* karena pada penelitian ini menggunakan bantuan *Software Mendeley Desktop*. Kemudian file yang di butuhkan akan di download secara otomatis. Adapaun hasil download tersebut menghasilkan data pada Tabel 1

Tabel 1 Contoh *Output Download Journal*

Label	*Zhang_2008_CS.DA_a*
Type	*JOUR*
Author	*Zhang, Chun-Xia and Zhang, Jiang-She*
Year	*2008*
Title	*A local boosting algorithm for solving classification problems*
Journal	*Computational Statistics & Data Analysis*

Tabel 1 Contoh *Output Download Journal* (lanjutan)

Volume	*52*
Issue	*4*
Pages	*1928-1941*
Abstract	*Based on the boosting-by-resampling version of Adaboost, a local boosting algorithm for dealing with classification tasks is proposed in this paper. Its main idea is that in each iteration, a local error is calculated for every training instance and a function of this local error is utilized to update the probability that the instance is selected to be part of next classifier's training set. When classifying a novel instance, the similarity information between it and each training instance is taken into account. Meanwhile, a parameter is introduced into the process of updating the probabilities assigned to training instances so that the algorithm can be more accurate than Adaboost. The experimental results on synthetic and several benchmark real-world data sets available from the UCI repository show that the proposed method improves the prediction accuracy and the robustness to classification noise of Adaboost. Furthermore, the diversity-accuracy patterns of the ensemble classifiers are investigated by kappa-error diagrams. © 2007 Elsevier B.V. All rights reserved.. . .*
Doi	*10.1016/j.csda.2007.06.015*
Keywords	*Adaboost; Classification noise; Kappa-error diagram; Local boosting; Weak learning algorithm*
ISSN	*0167-9473*
URL	*http//www.sciencedirect.com/science/article/pii/S0167947*

B. Preprocessing Data pada Term Perkata

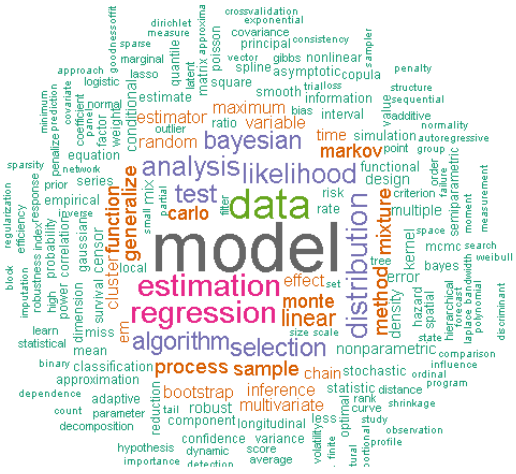
Berbagai ragam kata kunci yang telah dikumpulkan kemudian dilakukan praroses data dengan tahap *case folding, remove number, remove punctuation, remove stopwords* dan *verb, lemmatization,* dan *tokenizing*. Berikut dilakukan praproses data pada dokumen abstrak dari salah satu dokumen. Data hasil simulasi praproses ditunjukan sebagaimana pada Tabel 2.

Tabel 2 Praproses Data

keywords	*Association analysis and Biclustering and Biplots and Cluster analysis and Correspondence analysis and Data visualisation and Dimension reduction and Finite mixture and Fuzzy clustering and Multidimensional scaling*
Case Folding, Remove Number, & Remove Punctuation	*association analysis and biclustering and biplots and cluster analysis and correspondence analysis and data visualisation and dimension reduction and finite mixture and fuzzy clustering and multidimensional scaling*
Remove Stopwords	*association analysis biclustering biplots cluster analysis*

	correspondence analysis data visualisation dimension reduction finite mixture fuzzy clustering multidimensional scaling
Lemma-tization	*association analysis biclustering biplots cluster analysis correspondence analysis data visualisation dimension reduction finite mixture fuzzy cluster multidimensional scale*

Setelah dilakukan praproses data pada seluruh dokumen *keywords* maka proses selanjutnya yaitu membuat visualisasi *Word Cloud* untuk mengetahui kata yang paling sering muncul dari *keyword* dengan hasil yang bisa di lihat pada Gambar 9.

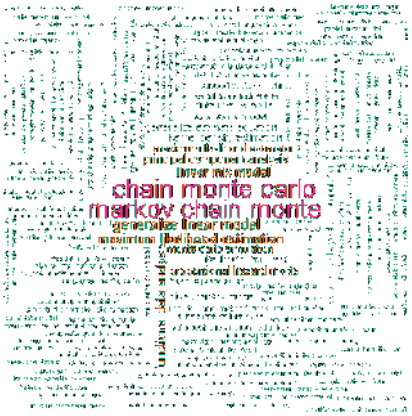


Gambar 9. *Word Cloud* data *keywords*

Pada Gambar 9 di atas dapat diketahui bahwa frekuensi kata terbanyak adalah *model*, yang diikuti dengan *data*, *distribution* dan kata setelahnya. Untuk kata *model*, *data*, dan *distribusi* pada diagram pareto ditampilkan frekuensi kemunculan kata berturut-turut sebesar 773, 429, dan 296 dengan frekuensi kumulatifnya berturut-turut yaitu sebesar 3,52%; 5,48%; dan 6,83% dari keseluruhan jumlah kata yang terlampir di buku. Kemudian di lanjutkan untuk membuat *wordcloud* bigram (dua kata) dengan visualisasi *wordcloud* pada Gambar 4.9.

Gambar 10. *Word Cloud* bigram

Pada wordcloud Gambar 10 dapat diketahui bahwa frekuensi kata terbanyak adalah *monte carlo,* kemudian di susul oleh *maximum likelihood, markov chain,* dan kata bigram setelahnya. Kemudian untuk nilainya sendiri yaitu frekuensi kemunculan berturut-turut pada kata *monte carlo, maximum likelihood, dan markov chain* berturut memiliki nilai 143, 111, dan 97 dan dengan persentase kumulatifnya berturut-turut adalah 0,67%; 1,2%; dan 1,65% dilihat dari diagram pareto yang terlampir di buku Kemudian di lanjutkan *wordcloud trigram* dengan visualisasi *wordcloud* pada Gambar 11.



Gambar 11. *Word Cloud* trigram

Gambar 11 menunjukkan bahwa frekuensi kata trigram terbanyak adalah *markov chain monte* dan *chain monte carlo,* dilanjutkan dengan kata selanjutnya. Untuk detail nilainya kata tersebut berturut-turut memiliki frekuensi kemunculan yang sama yaitu 83 dengan persentase satu satu dari keseluruhan sebesar 0,43% berdasarkan diagram pareto yang terlampir dalam buku. Kemudian untuk *wordcloud quadgram* dapat dilihat pada Gambar 12.

Gambar 12. *Word Cloud quadgram*

Gambar 12 menunjukkan bahwa *wordcloud quadgram* di dominasi oleh topik *markov chain monte carlo* di mana kata terebut merupakan kata yang paling menonjol di bandingkan dengan kata kata yang lain. Nilai frekuensi kemunculan yang di miliki kata *markov chain monte carlo* memiliki nilai sebanyak 83 kali kemunculan dengan persentase dari keseluruhan *quadgram* sebesar 0,48% dilihat dari digram pareto yang terlampir pada buku .

Kemudian dilanjutkan untuk menginput data menjadi data transaksi karena hendak di proses dengan *Association Rule*.

C. Association Rule pada Term Kata

Secara umum *association rule* mempunyai bentuk LHS dan RHS. Namun aturan tersebut di pakai jika menggunakan atribut *support* dan *confidence.* Namun pada kasus ini di pilih

frequent itemset karena pada penelitian ini LHS dan RHS pada transaksi tidak ada makna yang berbeda. Misal transaksi *regression => linear* sama saja dengan *linear => regression.* Sehingga hanya atribut *support* saja yang di gunakan.

Karena besarnya item yang dihasilkan maka untuk menentukan sebuah Topik bermakna maka digunakanlah c*losed frequent itemset* sebagai salah satu kriteria. c*losed frequent itemset* digunakan untuk mengurangi jumlah *frequent itemset* yang dihasilkan, sebagai contoh *association mining* dan *association rule mining* mungkin mengandung makna *itemset* yang sama. Dalam hal ini kita dapat menghapus salah satu *itemset* karena salah satunya merupakan *closed frequent itemset* dengan catatan memiliki jumlah frekuensi yang sama. Dan ada metode lebih lanjut dalam penghapusan *itemset* tersebut yaitu *remove subset* supaya menjadi lebih efisien dengan menghapus himpunan bagiannya. Bedanya dengan closed frequent itemset adalah subset rules tidak memandang frekuensi dari itemset tersebut. Maka untuk itu di lakukan perbandingan dari ketiga metode tersebut dengan nilai support yang berbeda sehingga menghasilkan jumlah *itemset* yang berbeda seperti yang dapat di lihat pada tabel 3.

Tabel 3 Perbandingan jumlah itemset pada tiga metode

min_support	Frequent itemset	Closed Frequent Itemset	Remove Subset
2 %	72	55	41
1 %	286	247	179
0,5 %	1.135	932	612
0,25 %	3.999	3009	1814
0,01 %	18.354.804	9.726.924	-

Tabel 3 merupakan perbandingan banyaknya itemset antara *frequent itemset, closed frequent itemset* dan *remove subset* dengan nilai *support* yang berbeda. Dari tabel tersebut dapat di lihat bahwa yang paling memiliki nilai terkecil dari perbandingan tiap *min support* adalah *remove subset* dari *frequent itemset.* Maka berikut hasil output dari remove subset dengan nilai *min support* 2,5% yang ditampilkan pada Tabel 4.

Tabel 4 Perbandingan Association Rule pada *min support* = 2%

No	***Itemset***	***Support***	***Frequent Itemset***	***Closed***	***Remove Subset***
1	data,model	0,073502	v	v	v
2	carlo,monte	0,066948	v	v	
3	linear,model	0,063202	v	v	
4	bayesian,model	0,057116	v	v	v
5	model,regression	0,056648	v	v	v
6	likelihood,maximum	0,055243	v	v	v
7	model,selection	0,049625	v	v	v
8	chain,markov	0,04588	v	v	
9	mixture,model	0,045412	v	v	v
10	generalize,model	0,044944	v	v	
⋮	⋮	⋮	⋮	⋮	⋮
72	maximum,model	0,020131	v	v	v

D. Correlation, Topic Network & Community

Setelah dilakukan *association rule* maka di lakukan korelasi antar tiap *itemset* untuk mencari pola hubungannya. Pada korelasi ini di gunakan data yang telah di lakukan remove subset karena data tersebut lebih mewakili himpunan *itemset* daripada *frequent itemset* dan *closed frequent itemset* pada support 2 %. Data yang dibandingkan adalah data banyaknya jurnal yang mengandung itemset tersebut di tiap tahunnya yaitu dari tahun 2008-2018 yang telah di proses pada rumus *excel* dan hasilnya bisa di lihat pada Tabel 4.

Tabel 4 menunjukkan ada sebanyak 41 variabel yang akan di lakukan korelasi berdasarkan frekuensi jurnal yang mengandung itemset pertahunnya di mana setelah di lakukan analisis korelasi menggunakan program R maka akan di hasilkan hasil output analisis *pearson correlation* yang dimana di dapatkan *p-value* dari tiap hubungan antar variabel. Maka hubungan variabel yang di gunakan adalah hanya pada yang *p-value* nya kurang dari nilai *alpha* yaitu 0,05 karena jika p<0,05 maka menandakan bahwa hubungan variabel tersebut berpengaruh secara signifikan. Sehingga akan di lakukan visualisai pada program *gephi* untuk membentuk *topic network* di mana variabel yang memiliki hubungan signifikan akan terbentuk garis. Visualisasi topic network tersebut dapat di lihat pada gambar 4.9 yang berada pada lembar selanjutnya.

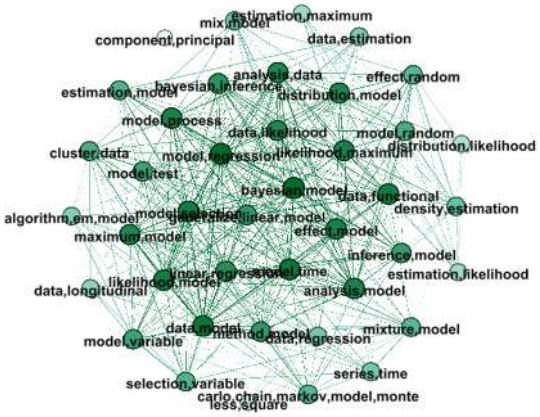
Tabel 5 Frekuensi jurnal yang mengandung itemset kata pertahunnya.

Var	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
x_1	7	0	3	18	17	16	32	10	16	20	18
x_2	6	0	7	13	22	10	19	8	13	12	12
x_3	6	0	5	17	15	15	17	8	11	14	13
x_4	10	0	5	17	14	16	15	8	17	8	8
x_5	5	0	6	12	13	13	18	6	10	8	15
x_6	7	0	8	11	8	3	22	7	16	9	6
x_7	5	0	1	14	15	8	15	8	11	3	11
x_8	6	0	5	17	13	13	12	5	8	5	6
x_9	4	0	5	9	12	13	9	5	10	10	13
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_{41}	3	0	1	5	5	8	3	5	8	2	3

Tabel 6. Keterangan nama itemset tiap kata.

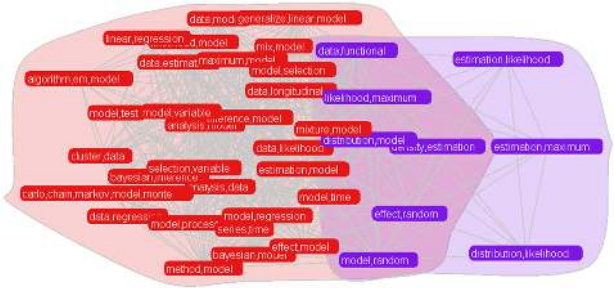
Var	Itemset	Var	Itemset
x_1	data,model	x_{22}	model,random
x_2	bayesian,model	x_{23}	method,model
x_3	model,regression	x_{24}	data,regression
x_4	likelihood,maximum	x_{25}	bayesian,inference
x_5	model,selection	x_{26}	data,likelihood
x_6	mixture,model	x_{27}	linear,regression
x_7	analysis,data	x_{28}	data,longitudinal
x_8	distribution,model	x_{29}	data,functional
x_9	estimation,model	x_{30}	model,time
x_{10}	likelihood,model	x_{31}	series,time
x_{11}	selection,variable	x_{32}	model,process
x_{12}	analysis,model	x_{33}	cluster,data
x_{13}	mix,model	x_{34}	effect,random
x_{14}	effect,model	x_{35}	component,principal
x_{15}	generalize,linear,model	x_{36}	less,square
x_{16}	model,test	x_{37}	inference,model
x_{17}	estimation,likelihood	x_{38}	algorithm,em,model
x_{18}	data,estimation	x_{39}	model,markov,chain,monte,carlo
x_{19}	model,variable	x_{40}	maximum,model
x_{20}	distribution,likelihood	x_{41}	estimation,maximum
x_{21}	density,estimation		

Kemudian keterangan nama itemset tiap kata per variabelnya dapat dilihat pada Tabel 6. Selanjutnya, topic network menggunakan Gephi dapat dilihat pada gambar 17.



Gambar 12. *Topic Network*

Gambar 12 di atas merupakan visualisasi hubungan pada tiap-tiap itemset yang di dapatkan dari *association rule* menggunakan *remove subset frequent itemset* yang saling terhubung melalui analisis korelasi sebelumnya. Terdapat sebanyak 41 *nodes* yang tersebar dan saling terhubung satu sama lain. Semakin besar nilai korelasinya maka garis penghubung antar itemset semakin besar. Begitu juga warna *node* yang semakin gelap menunjukkan bahwa *node* tersebut memiliki semakin banyak hubungan dengan itemset yang lainnya. Visualisasi tersebut menggunakan *software gephi* dengan *layout* penyebaran nodes menggunakan *Fruchterman Reingold.*dimana *layout* tersebut menggunakan gravitasi sehingga mampu membentuk seperti bola yang teratur. Untuk topic community bisa di lihat pada Gambar 18 menggunakan *clique percolation method* (CPM) pada software *CFinder.*



Gambar 13. *Topic Community*

Tabel 7 Analisis regresi pada 7 sampe variabel

tahun	Persamaan Regresi	R-sq	R-sq(adj)	Sparklines
x_1	-3152 + 1,573 tahun	34,3%	27,15%	
x_2	-1508 + 0,755 tahun	17,07%	7,85%	
x_3	-1691 + 0,845 tahun	25,53%	17,25%	
x_4	-740 + 0,373 tahun	4,99%	0%	
x_5	-1582 + 0,791 tahun	25,43%	17,15%	
x_6	-1107 + 0,555 tahun	9,46%	0%	
x_7	-1090 + 0,545 tahun	10,98%	1,08%	

Gambar 13 merupakan visualisasi *topic community* menggunakan CPM. Jadi garis-garis yang saling terhubung tersebut menunjukkan bahwa adanya hubungan yang signifikan dengan menggunakan *pearson correlation.* Kemudian terdapat dua sebaran komunitas itemset. Adanya irisan menunjukkan bahwa terdapat *overlaping* atau hubungan yang bertumpukan antar komunitas itemset atau topik. Untuk analisis regresinya adalah dengan variabel prediktornya yaitu tahun dan variabel responnya adalah frekuensi jurnal yang mengandung itemset tersebut dengan hasil outputnya yang bisa dilihat pada Tabel 7.

Dari Tabel 7 dapat diketahui persamaan regresi beedasarkan tiap tahunnya di atas. Pada nilai *R-square adsjusted* nilai tertingggi ada pada variabel x_1 yaitu "data,model" dimana nilai 27,15% tersebut menjelaskan berapa persen nilai pengaruh pergantian tahun yang di miliki terhadap hubungan *trend* kenaikan *itemset* tersebut.

E. *Preprocessing Data* pada *Term Keywords*

Berbagai ragam kata kunci yang telah dikumpulkan kemudian dilakukan praroses data dengan tahap *case folding,*saja karena *keywords* tersebut sudah menjadi bentuk term secara keseluruhan dalam *keywords* tersebut, sehingga tidak perlu melakukan penghapusan angka, tanda baca, kata kerja, dan *lemmatizatioin*. Berikut dilakukan praproses data pada dokumen abstrak dari salah satu dokumen. Data hasil simulasi praproses ditunjukan sebagaimana pada Tabel 8 berikut.

Tabel 8. Praproses Data pada *Term Keywords*

keywords	*Association analysis and Biclustering and Biplots and Cluster analysis and Correspondence analysis and Data visualisation and Dimension reduction and Finite mixture and Fuzzy clustering and Multidimensional scaling*
Case Folding, Remove Number, & Remove Punctuation	*association analysis and biclustering and biplots and cluster analysis and correspondence analysis and data visualisation and dimension reduction and finite mixture and fuzzy clustering and multidimensional scaling*

Kemudian kata *"and"* di atas merupakan pemisah tiap *term keywordsAssociation Rule* pada *Term Keywords.* Pada *term keywords* ini juga menggunakan *frequent itemset* menggunakan *remove subset* sebagai berikut.

Tabel 9. *Frequent itemset* pada *min support* = 0,15%.

Var	Itemset	support	n
x_1	missing data,multiple imputation	0.003280	7
x_2	asymptotic normality,consistency	0.003280	7
x_3	bayesian inference,markov chain monte carlo	0.002812	6
x_4	sensitivity,specificity	0.002812	6
x_5	em algorithm,mixture models	0.001874	4
x_6	dimension reduction,variable selection	0.001874	4
x_7	dimension reduction,sliced inverse regression	0.001874	4
x_8	birnbaumsaunders distribution,fatigue life distribution,lifetime data	0.001874	4
x_9	markov chain monte carlo,state space model	0.001874	4
x_{10}	aic,bic	0.001874	4
x_{11}	classification,regression trees	0.001874	4
x_{12}	kurtosis,skewness	0.001874	4
x_{13}	classification,linear discriminant analysis	0.001874	4
x_{14}	power,sample size	0.001874	4
x_{15}	generalized estimating equations,longitudinal data	0.001874	4
x_{16}	em algorithm,robust regression	0.001874	4
x_{17}	em algorithm,proportional hazards model	0.001874	4
x_{18}	outliers,robust estimation	0.001874	4
x_{19}	mixture models,model-based clustering	0.001874	4
x_{20}	gibbs sampling,markov chain monte carlo	0.001874	4

Var	Itemset	support	n
x_{21}	quantile regression,variable selection	0.001874	4
x_{22}	em algorithm,random effects	0.001874	4
x_{23}	bootstrap,simulation	0.001874	4
x_{24}	bayesian inference,mcmc	0.001874	4

Frekuensi tiap tahunnya dapat dilihat pada Tabel 9.

Tabel 9. *Frequent itemset* pada *min support* = 0,15%.

Var	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
x_1	0	0	2	0	1	0	1	1	1	1	0
x_2	0	0	0	3	1	1	1	0	1	0	0
x_3	1	0	0	2	1	1	2	0	0	0	0
x_4	0	0	0	0	2	0	1	1	0	1	1
x_5	0	0	0	1	0	0	2	0	1	1	1
x_6	0	0	0	0	0	2	1	1	0	1	1
x_7	0	0	0	1	0	2	1	0	0	0	0
x_8	0	0	0	3	0	1	0	0	0	0	0
x_9	0	0	0	0	2	0	2	0	0	0	0
x_{10}	1	0	0	0	0	2	0	1	0	0	0
x_{11}	0	0	1	0	1	0	1	0	0	1	0
x_{12}	2	0	0	1	1	0	0	0	0	0	0
x_{13}	0	0	0	0	0	0	1	0	1	0	2
x_{14}	0	0	0	2	0	1	1	0	0	0	0
x_{15}	0	0	0	0	0	1	1	1	0	1	0
x_{16}	0	0	0	0	2	0	1	0	0	1	0
x_{17}	0	0	0	0	0	1	0	1	1	0	1
x_{18}	0	0	0	0	0	1	0	1	0	2	0
x_{19}	0	0	0	0	0	0	2	0	2	0	0
x_{20}	1	0	0	1	0	1	1	0	0	0	0
x_{21}	0	0	0	0	0	0	2	0	1	0	1
x_{22}	1	0	1	0	0	0	1	0	1	0	0
x_{23}	2	0	0	1	0	0	0	0	0	1	0
x_{24}	1	0	0	0	1	0	0	0	2	0	0

Pada Tabel 9 dapat dilihat bahwa itemset *missing data,multiple imputation* dan *asymptotic normality,consistency* merupakan itemset terbanyak dengan nilai support 0,003280225 dan jumlah itemset masing-masing sebanyak 7.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan analisis dan pembahasan yang telah dilakukan pada bab 4, maka diperoleh kesimpulan sebagai berikut

1. Pada saat setelah di lakukan preprocessing term kata pada *keywords* kemudian di visualisasikan dengan *wordcloud* maka kata-kata yang paling sering muncul adalah *model*, *data*, dan *distribution.* Pada worcloud bigram, dua kata yang paling sering muncul adalah *monte carlo, maximum likelihood,* dan *markov chain.* Sedangkan pada wordcloud trigram tiga kata yang paling sering muncul adalah *markov chain monte, chain monte carlo,* dan *generalize linear model*. Sedangkan wordcloud quadgram didominasi oleh *markov chain monte carlo*.
2. Pada *association rule term* kata jika menggunakan min support 2 % maka di dapatkan *frequent itemset* sebanyak 72, *closed frequent itemset* sebanyak 55, dan *remove subset* sebanyak 41 itemset. Kemudian saat di lakukan analisis korelasi pada *itemset remove subset.* Di dapatkan *bayesian,model* yaitu itemset yang paling banyak memiliki hubungan. Kemudian pada *topic community* dengan *cfinder* terbagi menjadi dua komunitas dan terdapat irisan sebanyak 6 *itemset*.
3. Pada *association rule term keywords* jika menggunakan apriori dengan min support 1,5 % maka di dapatkan sebanyak 27 itemset yang terlampir.
4. Metode *association rule* ini mampu mencari kandidat topik dari sekumpulan kata yang dalam penelitian ini adalah gabungan dari banyak keyword yang di analisa, namun dengan catatan topik tersebut terdiri dari lebih dari satu item, karena metode tersebut mencari kata yg sering berhubungan dan topik seringkali bersifat demikian, seperti kata *markov* yang senantiasa berdampingan dengan kata *chain*, *monte*, dan *carlo.*

B. Saran

Berdasarkan kesimpulan yang diperoleh, dapat dirumuskan saran sebagai pertimbangan penelitian selanjutnya adalah pengoptimalan tema data pada jurnal penelitian yang di ambil untuk memudahkan dalam pemahaman terkait topik yang ada dalam jurnal tersebut serta di butuh *analysis platform* yang mumpuni guna meringankan beban komputasi untuk analisis.

DAFTAR PUSTAKA

[1] L. H. Jose, A. Ankur and Z. Xingquan, Topic discovery and future trend forecasting for texts, Florida: Department of Computer and Electrical Engineering and Computer Science Florida Atlantic University, 2016.

[2] U. Setyawati, 2008. [Online]. Available: https://elib.unikom.ac.id/gdl.php?mod=browse.

[3] Elsevier, 2015. [Online]. Available: http://digilib.undip.ac.id/v2/wpcontent/uploads/2016/. [Accessed 23 January 2019].

[4] D. M. Christopher, P. R. and S. Hinrich, Introduction to Information Retrieval 1st Edition, England: Cambridge University Press, 2008.

[5] Z. Chengqi, Association Rule Mining : Models and Algorithms, 2002.

[6] L. E. T. Kusrini, Algoritma Data Mining, Yogyakarta.: Andi, 2009.

[7] D. Kusumo, M. Bijaksana and D. Darmantoro, "Data Mining Dengan Algoritma Apriori Pada RDBMS Oracle," *Jurnal Penelitian,* p. 1–5, 2016.

[8] Priyanto and D., Mandiri Belajar Analisis Data Dengan SPSS, Yogyakarta: Mediakom, 2013.

[9] N. Pasquier, Y. Bastide, R. Taoull and L. Lakhal, "Efficient Mining of Association Rules Using Closed Itemset Lattices," *Information Systems,* 1999.

[10] L. Zhuang and H. Dai, "A Maximal Frequent Itemset Approach for Web Document Clustering," in *4th International Conference on Computer and Information Technology*, 2004.

[11] R. Walpole, Intoduction to Statistics, New York: Macmillan Publishing Co. Inc, 1995.