TEXT MINING MENGGUNAKAN GENERATE ASSOCIATION RULE WITH WEIGHT (GARW) ALGORITHM UNTUK ANALISIS TEKS WEB CRAWLER

Zulkifli Arsyad
Politeknik Negeri Bandung
zulkifli.arsyad@jtk.polban.ac.id

ABSTRACT

Text mining is widely used to find hidden patterns and information in a large number of semi and unstructured texts. Text mining extracts interesting patterns to explore knowledge from textual data sources. Association rule extraction GARW (Generating Association Rule using Weighting Scheme) can be used to find knowledge from a collection of web content without having to read all the web content manually from the many search results of crawlers. The GARW algorithm is a development of a priori to produce relevant association rules. From the results of this knowledge discovery can facilitate netizens users in finding relevant information from search keywords without having to review one by one web content generated from search engine searches.

Keyword: *Text Mining, Association Rule, GARW,Web Crawler.*

ABSTRAK

Text mining secara luas digunakan untuk menemukan pola yang tersembunyi dan informasi dalam sejumlah besar teks semi dan tidak terstruktur. *Text mining* mengekstrak pola yang menarik untuk mengeksplorasi pengetahuan dari sumber data tekstual. *Association rule extraction GARW (Generating Association Rule using Weighting Scheme)* dapat digunakan untuk menemukan pengetahuan dari kumpulan konten web tanpa harus membaca semua konten *web* secara manual dari sekian banyaknya hasil pencarian *web crawler search engine.* Algoritma GARW merupakan pengembangan dari apriori untuk menghasilkan aturan asosiasi yang relevan. Dari hasil penemuan pengetahuan ini dapat memudahkan pengguna netizen dalam menemukan informasi yang relevan dari kata kunci pencarian tanpa harus mereview satu persatu konten web yang dihasilkan dari pencarian *search engine.*

Kata Kunci: *Text Mining, Association Rule, GARW,Web Crawler.*

1. PENDAHULUAN

Proses menganalisa data dan informasi dari internet, menemukan rangkuman informasi yang tepat dari berbagai halaman web dengan teks yang semi terstruktur serta tidak terstruktur saat ini dilakukan secara manual, kesempatan untuk mengabaikan data yang tidak terstruktur merupakan hal yang penting dalam dunia kompetitif saat ini, menurut survey IDC, data tidak terstuktur menempati 80% dibandingkan dengan hanya 20% untuk data terstruktur. Hal ini menggambarkan bahwa diperlukan waktu untuk memilah data yang diterima dari 80% data yang tidak terstruktur tersebut.

Teks mining merupakan teknik untuk menganalisa sejumlah besar teks bahasa alami, dan mendeteksi pola leksikal untuk mengekstrak informasi yang berguna, *text mining* ini digunakan untuk menemukan informasi yang menarik dari basis data yang sangat besar (R & Vijayarani, 2016).Secara konsepnya *teks mining* merupakan teknik yang digunakan untuk menangani permasalahan klasifikasi, kluster, ekstraksi informasi dan pencarian informasi (R. Talib, 2016). *Text mining* secara luas digunakan untuk menemukan pola yang tersembunyi dan informasi dalam sejumlah besar teks semi serta tidak terstruktur (Naithani, 2016). Hal ini menjadi *open issue* yang menarik untuk di analisis bagaimana melakukan ekstraksi informasi

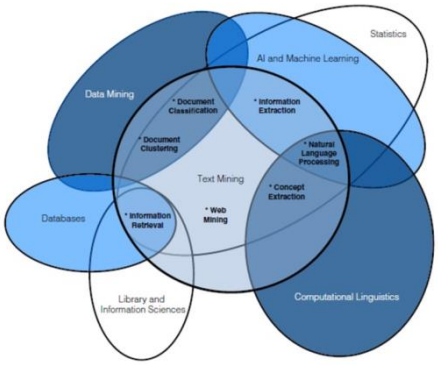
dari data yang tidak terstruktur pada web dari data yang tersebar di berbagai server di seluruh dunia.

Text mining atau penemuan pengetahuan merupakan sub proses dari data mining, yang secara luas digunakan untuk menemukan pola yang tersembunyi dan informasi dalam sejumlah besar teks tidak terstruktur (Naithani, 2016). T*ext mining* mengekstrak pola yang menarik untuk mengeksplorasi pengetahuan dari sumber data tekstual, sumber yang menjadi kajian peneliti ini bersumber dari web yang menghasilkan banyak sekali konten tekstual yang beraneka ragam strukturnya (R. Talib, 2016). *Text mining* merupakan pendekatan multi-disiplin atau pendekatan yang berdasarkan pada pencarian informasi, *data mining*, *machine learning*, statistik, dan komputasi linguistik (R. Talib, 2016).

Dalam *text mining* terdapat tahapan-tahapan yang umumnya dilakukan (R & Vijayarani, 2016).

1. Mengkonversi data yang tidak terstruktur menjadi terstuktur,
2. Mengidentifikasi pola dari struktur data,
3. Menganalisis pola menggunakan algoritma *text mining*,
4. Mengekstrak informasi yang bermanfaat dari text.

Untuk menjalankan setiap tahapan dalam *teks mining*, terdapat teknik atau pendekatan yang dapat dipilih yang disesuaikan dengan domain yang akan diteliti, berikut ini gambar interaksi *text mining* antar pendekatan (R. Talib, 2016).



Gambar 1. Diagram interaksi text mining antar pendekatan

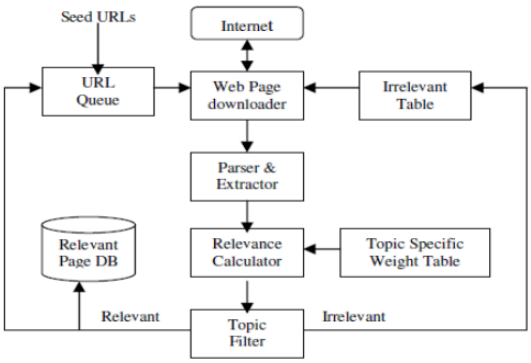
Pada diagram tersebut terlihat keterkaitan antar pendekatan yang terdapat pada *text mining* ini, peneliti menentukan web mining yang menjadi issue yang cukup menarik untuk dikaji lebih jauh mengingat, konten yang sangat banyak dari web yang dihasilkan serta struktur yang berbeda-beda , dan memiliki interaksi dengan beberapa pendekatan dalam *text mining.*

Dalam pencarian kata kunci pada *search engine*, terdapat banyak sekali tautan halaman web yang dimunculkan pada *search engine*. Dari laman yang ditampilkan pada hasil pencarian dapat kita lakukan analisis untuk menemukan rangkuman informasi yang dibutuhkan dari masing-masing laman yang di tampilkan berdasarkan pencarian kata kunci. Konsep tersebut menjadi menarik jika dilengkapi dengan *association rule extraction* yang sangat berguna untuk menemukan pengetahuan dari kumpulan dokumen tanpa membaca semua dokumen secara manual (Kulkarni, 2016). Dalam *association rule* ini dihasilkan dari aturan yang memenuhi kriteria seperti *support*, *confidence*, dan nilai TF-IDF kata kunci. Algoritma ini disebut sebagai GARW (*Generating Association Rule using Weighting Scheme*). Algoritma GARW beroperasi dengan cara yang sama seperti algoritma apriori akan tetapi ditambahkan dengan beberapa langkah tambahan untuk menyelesaikan masalah apriori dan untuk menghasilkan aturan asosiasi yang relevan (Janwe, 2011). Algoritma GARW digunakan dalam proses *web crawler* dalam menemukan pengetahuan yang berguna dari hasil pencarian. *Web crawler* adalah sistem untuk mengunduh sebagian besar halaman web, pada pengelolaan informasinya web *crawler* tidak terpusat pada *repository* atau *datasource* tertentu, melainkan terdiri dari ratusan juta

penyedia konten web yang independen, masing-masing menyediakan layanan mereka sendiri (Christopher Olston, 2010). Berdasarkan pemaparan tersebut peneliti akan membuat analisis *text mining* menggunakan algoritma *association rule extraction* GARW untuk menemukan pengetahuan dari hasil pencarian web crawler pada *search engine*, hasil dari pencarian *search engine* , memiliki banyak sekali tautan konten web akan di ekstrak untuk diambil informasi yang relevan berdasarkan kata kunci yang diinputkan.

2. METODOLOGI

Web crawler (juga dikenal sebagai robot web atau spider) adalah sebuah program untuk men-*download* Halaman web. Melakukan *download* halaman yang sesuai, ekstrak semua URL yang terkandung di dalamnya, dan menambahkan setiap URL yang sebelumnya tidak diketahui. *Web crawler* mencoba untuk mencari dan mengambil halaman web yang relevan dengan domain yang spesifik (Najork & Heydon, 2001).

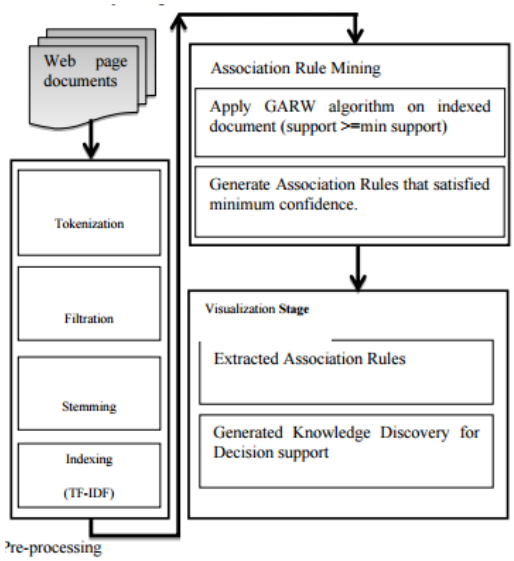


Gambar 2. Arsitektur web crawler

Pada gambar diatas *web crawler* fokus mengambil konten dari halaman yang relevan yang secara bersamaan melintasi sejumlah minimal dokumen / list pencarian yang tidak atau kurang relevan di web. Dalam *web crawler*, *seed* URL digunakan untuk meninisialisasi proses *crawling*. Setiap URL yang dikunjungi oleh *web crawler* ini mengidentifikasi *hyperlink* yang berbeda di halaman yang berbeda dari internet oleh *parser* dan *generator* yang disimpan dalam sistem basis data mesin pencari (Kulkarni, 2016).

Berikut adalah gambar *Generate Association Rule with Weight* (GARW) untuk membangkitkan penemuan pengetahuan secara umum dilakukan melalui tiga tahapan (Kulkarni, 2016) :

1. *Text Pre-processing*
2. *Association rule mining*
3. *Visualization stage*



Gambar 3. Blok diagram untuk penemuan pengetahuan menggunakan association rule

Keterangan :

1. *Tokenization*
   Tokenisasi adalah proses memisahkan teks ke dalam kata-kata atau istilah. Fase ini memainkan peran yang sangat penting dalam menghasilkan aturan asosiasi. Tujuan utamanya adalah untuk mengkonversi dokumen tidak terstruktur ke dalam dokumen terstruktur. Halaman web biasanya berisi informasi dalam format yang tidak terstruktur. ini membuat masalah bagi pertambangan teks yang akurat dan relevan. Oleh karena itu diperlukan untuk mengkonversi informasi tidak terstruktur untuk format yang terstruktur.
2. *Filtration*
   Mencari informasi yang diharapkan memiliki output yang relevan dari pencarian. Dalam pencarian informasi terdapat pesoalan bahwa input script mengandung terlalu banyak kata-kata atau tag yang beragam dengan struktur yang berbeda. Untuk menghasilkan pengetahuan yang akurat dari koleksi halaman web, pengguna perlu untuk mengetahui hubungan antara semua kata kunci namun peran ini dapat menjadi tidak efisien jika kita tidak menghapus kata-kata atau scipt yang yang tidak dibutuhkan pada dokumen input. pada tahapan ini hasil pencarian pada halaman web yang dapat berupa kalimat, paragraf atau dokumen dibagi menjadi bagian-bagian tertentu kemudian token disaring dengan menghapus *stop words* dan sufiks untuk mencapai hasil penemuan pengetahuan yang lebih baik. Adapun *stop words* yang harus di hapus adalah seperti pada tabel 1.

Tabel 1. Stop words Bahasa Indonesia dan Inggris

***Stop words* Bahasa Inggris**	***Stop words* Bahasa Indonesia (Tala, 2003).**
a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, i, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no,	ada, adalah, adanya, adapun, agak, agaknya, agar, akan, akankah, akhir, akhiri, akhirnya, aku, akulah, amat, amatlah, anda, andalah, antar, antara, antaranya, apa, apaan, apabila, apakah, apalagi, apatah, artinya, asal, asalkan, atas, atau, ataukah, ataupun, awal, awalnya, bagai, bagaikan, bagaimana, bagaimanakah, bagaimanapun, bagi, bagian, bahkan, bahwa, bahwasanya,

nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your	baik, bakal, bakalan, balik, banyak, bapak, baru, bawah, beberapa, begini, beginian, beginikah, beginilah, begitu, begitukah, begitulah, begitupun, bekerja, belakang, belakangan, belum, belumlah, benar, benarkah, benarlah, berada, berakhir, berakhirlah, berakhirnya, berapa, berapakah, berapalah, berapapun, berarti, berawal, berbagai, berdatangan, beri, berikan, berikut, berikutnya, berjumlah, berkali-kali, dll.

3. Stemming

Adalah sebuah proses di mana bentuk-bentuk varian dari kata yang sama dikurangi sehingga menjadi bentuk yg umum. Hal ini penting untuk memahami bahwa kita menggunakan *stemming* ini dengan tujuan untuk meningkatkan kinerja sistem penemuan pengetahuan. Algoritma porter adalah proses untuk menghapus suffix dari kata-kata dalam bahasa inggris. Pada table 2 berikut menunjukan beberapa peraturan yang dimodifikasi menggunakaan algoritma porter.

Table 2. Modifikasi pada Aturan Porter Stemming Bahasa Inggris

No	Rule added/Changed	Example	Output by original Porter	Output by modified Porter
1	OUS -> OUS	Obvious	Obviou	Obvious
2	ICAL -> ICAL	Medical	Medice	Medical
3	IAL ->E	Official	Offici	Office
4	ANT->ANT	Servant	Serv	Servant
5	DENT-DENT	Accident	Accid	Accident
6	ATION-> ATE	Corporat ion	Corpor	Corporate
7	E -> E	Village	Villag	Village
After removal of "ing" character use replacement rule:				
8	S -> SE	Exposing	Expos	Expose
After removal of „ed‟ character use replacement rule:				
9	R -> RE	Declared	Declar	Declare
10	D -> DE	Provided	Provid	Provide
11	G -> GE	Emerged	Emerg	Emerge
12	V -> VE	Removed	Remov	Remove
13	FI -> FY	Identified	Identifi	Identify

Adapun tahapan dalam proses stemming dalam bahasa indonesia adalah sebagai berikut :

a. Hapus *Particle*
b. Hapus *Possesive Pronoun*
c. Hapus awalan pertama. Jika tidak ada lanjutkan ke langkah 4a,jika ada cari maka lanjutkan ke langkah 4b
d. i). Hapus awalan kedua, lanjutkan ke langkah 5a
   ii). Hapus akhiran, jika tidak ditemukan maka kata tersebut diasumsikan sebagai *root word*. Jika ditemukan maka lanjutkan ke langkah 5b.
e. i). Hapus akhiran. Kemudian kata akhir diasumsikan sebagai *root word*

ii). Hapus awalan kedua. Kemudian kata akhir diasumsikan sebagai *root word.*

4. *Indexing*

Skema pembobotan TF-IDF (*Term Frequency*, *Inverse Document Frequency*) digunakan untuk menetapkan bobot untuk membedakan istilah dalam dokumen. Frekuensi istilah hitungan yang mewakili berapa kali kata kunci (x) telah terjadi dalam dokumen. Frekuensi dokumen invers adalah hitungan yang mewakili jumlah total dokumen yang berisi kata kunci (x) setidaknya sekali. Kami telah menggunakan metode TF-IDF untuk mendapatkan penemuan pengetahuan yang relevan hanya untuk menghitung total kata kunci dalam dokumen masukan. Adapun rumus untuk menentukan TF adalah (Kulkarni, 2016):

$$(tf)_{i,j} = \sum_{i=j=1}^{n} (Nt_1, d_{nj}) \qquad (1)$$

Diketahui :

Nt_i , d= Jumlah istilah t_i yang terjadi di dalam dokumen d_j

Inverse Document Frequency (Kulkarni, 2016)

$$(idf)_{i,j} = \sum_{i=j=1}^{n} log\left(\frac{|C|}{Nt_i}\right) \qquad (2)$$

Diketahui :

Nt_i menunjukan jumlah dokumen dalam koleksi C, dimana t_i terjadi setidaknya sekali.dan |C| menunjukan total dari dokumen.

Vector Space Model – Cosine Similarity

Formally, a vector space is defined by a set of linearly independent basis vectors
Setelah memiliki bobot dengan istilah yang berbeda menggunakan TF / IDF, pengecekan kesamaan antara query dan isi pada dokumen web diperlukan untuk menghitung ambang batas threshold berdasarkan kesamaan istilah (Bathla & Jindal, 2011). Model Ruang Vektor digunakan untuk menghitung kesamaan ini (Bathla & Jindal, 2011):

$$|Di| = \sqrt{wd^2} \qquad (3)$$

$$|Q| = \sqrt{wq^2} \qquad (4)$$

$$\text{Cosin e}\emptyset di = \frac{Q.Di}{|Q|.|Di|} \qquad (5)$$

$$\text{Sim(Q,Di)} = \text{Cosin e}\emptyset di \qquad (6)$$

Dimana Q.Di adalah dot produk dari bobot query .
|Q| merupakan panjang dari query vector
|Di| merupakan panjang dari dokumen vector

Association Rule terdiri dari struktur IF-ELSE yang dapat memprediksi atribut kombinasi. Untuk setiap konsekuen aturan IF THEN kita menghitung *support* dan *confidence* yang cocok dengan nilai-nilai yang ditentukan (Bathla & Jindal, 2011). Dalam *association rule* ini dihasilkan dari association rule yang memenuhi kriteria seperti *support*, *confidence*, dan nilai TF-IDF kata kunci. Algoritma ini disebut sebagai GARW (*Generating Association Rule using Weighting Scheme*). Algoritma GARW beroperasi dengan cara yang sama seperti algoritma Apriori akan tetapi ditambahkan dengan beberapa langkah tambahan untuk menyelesaikan masalah Apriori dan untuk menghasilkan aturan asosiasi yang relevan (Bhujade & Janwe, 2011). Adapun Algoritma GARW secara deskriptif dijelaskan sebagai berikut (Kulkarni, 2016):

1. Pindai file yang berisi semua kata kunci yang memenuhi nilai ambang batas (*threshold* weight) dan frekuensi masing-masing dokumen.

2. "N" menunjukkan jumlah kata kunci atas yang memenuhi nilai bobot *threshold*,
3. Letakkan paling atas kata kunci N dalam daftar indeks bersama dengan nilai TFdi dalam semua daftar,
4. Gunakan minimum *support* dan *confidence* untuk mengenerate *association rule*,
5. Scan list index dan temukan semua kata kunci yang memenuhi ambang batas dukungan minimal,
6. Dalam K >=2, kandidat kata kunci C_k ,dimana nilai k di hasilkan dari set kata kunci.
7. Pindai daftar indeks dan frekuensi menghitung dari set kata kunci C_k yang dihasilkan pada langkah 5,
8. Bandingkan frekuensi set calon kata kunci dengan minimum support,
9. Set kata kunci k yang memenuhi minimum support, temukan dari langkah 6,
10. Untuk setiap set kata kunci yang sering muncul , temukan semua asosiasi aturan yang memenuhi ambang minimum *confidence*.

3. HASIL DAN PEMBAHASAN

3.1 Business Understanding

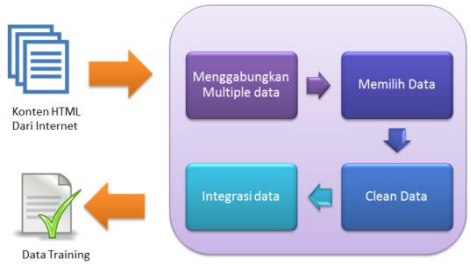
Pada tahapan ini difokuskan pada memahami kebutuhan dan tujuan peneliti dari sudut pandang bisnis, pada pencarian di internet pengguna perlu mencari tahu hubungan antara semua kata kunci yang ada di halaman web halaman tersebut, bagi pengguna membaca kumpulan dokumen dan mendapatkan beberapa pengetahuan memakan waktu dan kurang efektif. Pada penelitian ini peneliti akan mempermudah proses pencarian dengan mengambil konten konten tertentu yang perlu berdasarkan pemilihan kata kunci dan kategori .

3.2 Data Understanding

Sumber data berasal dari konten halaman web dari hasil pencarian *search engine google*, data dari konten halaman web tersebut yang berupa dokumen tag html akan dilakukan ekstraksi untuk mendapatkan konten artikel utama pada laman tersebut. Konten yang dihasilkan dari hasil pencarian berdasarkan kata kunci ini awalnya masih berupa konten HTML yang memiliki struktur yang bervariasi yang merupakan data semi terstruktur, multi bahasa, serta banyak tag HTML yang tidak diperlukan yang perlu untuk di bersihkan.

3.3 Data Preparation

Data dari internet perlu dilakukan pembersihan data dari konten yang tidak perlu secara otomatis. Mekanisme pembersihan data adalah sebagai berikut :



Gambar 4. Tahapan dalam Data Preparation

Keterangan :

1. Menggabungkan Multiple Data
   Konten dari internet dapat berasal dari laman yang memiliki struktur HTML berbeda, dan dari sumber data yang berbeda-beda dari jutaan server di dunia. Untuk bahan analisa peneliti mengambil sample dari hasil pencarian yang didapat beberapa link URL dan konten berikut , dalam kasus ini kata kunci yang digunakan sebagai bahan untuk analisis adalah "text mining" , dengan total pencarian konten di internet penulis membatasi 100

laman konten web, namun dalam pembahasan analisis dalam menemukan pengetahauan konten yang di jelaskan adalah 5 laman konten web, mewakili semua laman hasil pencarian.

Tabel 3. Data Laman Konten Web

Alamat URL	Konten
http://searchsqlserver.techtarget.com/definition/data-mining	<html class="gt-ie8 js no-touch csstransforms csstransforms3d csstransitions mti-inactive" dir="ltr" lang="en" prefix="og: http://ogp.me/ns#" itemscope="" itemtype="http://schema.org/Article" style=""><!--<![endif]--><head><script src="http://pagead2.googlesyndication.com/pagead/osd.js"></script><script src="https://securepubads.g.doubleclick.net/gampad/ads?gdfp_req=1&correlator=779309671201869&output=json_html&callback=googletag.impl.pubads.callbackProxy5&impl=fif&eid=108809107%2C21060362%2C21060364&sc=0&sfv=1-0-.... to predict future trends. ….
https://en.wikipedia.org/wiki/Data_mining	<html class="client-js ve-not-available" lang="en" dir="ltr"><head><meta charset="UTF-8"><title>Data mining - Wikipedia</title><script>document.documentElement.className = document.documentElement.className.replace(/(^|\s)client-nojs(\s|$)/, "$1client-js$2");</script><script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgCanonicalNames … …
https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm	<html lang="en" head> aid="CHDDIABC"><h1 class="chapter">1 What Is Data Mining?</h1> This chapter provides a high-level orientation to data mining technology. Note: Information about data mining is widely available. No matter what your level of expertise, you will be able to find helpful books and articles on data mining. Here are two web sites to help you get started: <code>http://www.kdnuggets.com/</code> — This site is an excellent source of information about data mining. It includes a bibliography of publications. …
https://id.wikipedia.org/wiki/Penggalian_data	<html class="client-js ve-available" lang="id" dir="ltr"><head><meta charset="UTF-8"><title>Penggalian data - Wikipedia bahasa Indonesia, ensiklopedia bebas</title><script>document.documentElement.className = document.documentElement.className.replace(/(^|\s)client-nojs(\s|$)/, "$1client-js$2");</script>

	<script>(window.RLQ=window.RLQ||[]).push(function(){mw.config.set({"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":false,"wgNamespaceNumber":0,"wgPageName":"Penggalian_data",….< a href="#cite_note-1">[1]</sup>. Suatu pola dikatakan menarik apabila pola tersebut tidak sepele, implisit, tidak diketahui sebelumnya, dan berguna. Pola yang disajikan haruslah mudah dipahami, berlaku untuk data yang akan diprediksi dengan derajat kepastian tertentu, berguna, dan baru. Penggalian data memiliki beberapa nama alternatif, meskipun definisi eksaknya berbeda, seperti KDD (knowledge discovery in database), analisis pola, arkeolog…
http://www.thearling.com/text/dmwhite/dmwhite.htm	<html><head> <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1"> <meta name="description" content="An Introduction to Data Mining"> <meta name="keywords" content="consult, consulting, consultant, data mining, database marketing, database mining, data analysis, dss, decision support, olap, kdd, kdnuggets, knowledge discovery, data visualization, information design, information visualization, one to one marketing, one-to-one marketing, marketing, kurt thearling, thearling"> <meta name="GENERATOR" content="Microsoft FrontPage 4.0"> <title>An Introduction to Data Mining</title> .. <p>Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automate …

2. Memilih Data
   Konten yang diambil dari halaman web tersebut merupakan konten yang terdapat pada tag title, pageTitle, content, bodyContent, pageContent, content-center, article dari masing-masing laman yang didapatkan dari hasil pencarian. Id Tag div yang di tentukan penulis berdasarkan hasil ektraksi dari beberapa url , didapat beberapa penamaan tag div dengan id konten yang beragam, namun peneliti mencoba mengidentifikasi penamaan id konten yang umumnya digunakan dalam laman web.
3. Membersihkan Data
   Tag tag pada konten HTML yang tidak dibutuhkan akan dihilangkan.hanya konten yang terdapat dalam konten utama yang akan di pertahankan.dari table 5.1 , tag html yang tidak diperlukan akan dihilangkan, kemudian tag yang berada dalam id content akan diambil sebagai *data training*, berikut hasil pembersihan data yang ditampilkan dalam table 4.2.

Tabel 4. Konten data training yang telah dibersihkan

Alamat URL	Konten

http://searchsqlserver.techtarget.com/definition/data-mining	Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends. Download this free guide Download Our Exclusive Big Data Analytics Guide An unbiased look at real-life analytics success stories, including a Time Warner Cable case study, and tips on how to evaluate big data tools. This guide will benefit BI and analytics pros, data scientists, business execs and project managers. Start Download Corporate E-mail Address: You forgot to provide an Email Address. This email address doesn't appear to be valid. This email address is already registered. Please login. You have exceeded the maximum character limit. Please provide a Corporate E-mail Address. ….
https://en.wikipedia.org/wiki/Data_mining	Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.[1] It is an interdisciplinary subfield of computer science.[1][2][3] The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.[1] Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data …
https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm	Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD). The key properties of data mining are: Automatic discovery of patterns Prediction of likely outcomes Creation of actionable information Focus on large data sets and databases Data mining can answer questions that cannot be addressed through simple query and reporting techniques. Automatic Discovery Data mining is accomplished by building models. A model uses an algorithm to act on a set of data. The notion of automatic discovery refers to the execution of data mining models.
https://id.wikipedia.org/wiki/Penggalian_data	Penggalian data (bahasa Inggris: *data mining*) adalah ekstraksi pola yang menarik dari data dalam jumlah besar [1]. Suatu pola dikatakan menarik apabila pola tersebut tidak sepele, implisit, tidak diketahui sebelumnya,

	dan berguna. Pola yang disajikan haruslah mudah dipahami, berlaku untuk data yang akan diprediksi dengan derajat kepastian tertentu, berguna, dan baru. Penggalian data memiliki beberapa nama alternatif, meskipun definisi eksaknya berbeda, seperti KDD (knowledge discovery in database), analisis pola, arkeologi data, pemanenan informasi, dan intelegensia bisnis. Penggalian data diperlukan saat data yang tersedia terlalu banyak (misalnya data yang diperoleh dari sistem basis dataperusahaan, e-commerce, data saham, dan data bioinformatika), tetapi tidak tahu pola apa yang bisa didapatkan. …
http://www.thearling.com/text/dmwhite/dmwhite.htm	An Introduction to Data Mining Discovering hidden value in your data warehouse Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

4. Integrasi Data

   Konten data *training*, kemudian disimpan ke dalam basis data penyimpanan sementara *(Array Map*) atau database sebagai data *training*. *Array Map* seperti pada dekalrasi map dibawah ini.

   Map<String, String> map = new HashMap<String, String>();

   Merupakan variable yang bertipe string.yang berisi data training yang telah di bersihkan. Data yang tersimpan baik dalam Array Map atau database telah distrukturkan sehingga memudahkan dalam proses preprocessing selanjutnya.

5. *Tokenization*

   Pada tahapan tokenisasi ini memisahkan teks dalam kalimat menjadi beberapa bagian, dalam data yang didapat dari hasil pencarian , setiap kata di bagi seperti pada gambaran berikut :

Tabel 5. *Tokenization*

No	Alamat URL	Konten
1	http://searchsqlserver.techtarget.com/definition/data-mining	Data|mining|is|the|process|of|sorting|through|large|data|sets|to|identify|patterns|and|establish|relationships|to|solve|problems|through|data|analysis.|Data|mining|tools|allow|enterprises|to|predict|future|trends.|Download|this|free|guide|Download|Our|Exclusive|Big|Data|Analytics|Guide|An|unbiasedlook|at|real-life|analytics|success|stories,|

		including|a|time|Warner|Cable|case|study|and|tips|on|how|to|evaluate|big|data|tools.|
2	https://en.wikipedia.org/wiki/Data_mining	Data|mining|is|the|computing|process|of|discovering|patterns|in|large|data|sets|involving|methods|at|the|intersection|of|machine|learning,|statistics,|and|database|systems.[1]|It|is|an|interdisciplinary|subfield|of|computer|science|The|overall|goal|of |the|data|mining|process|is|to|extract|information|from|a|data| set|and|transform|it|into|an|understandable|structure|for |further|use
3	https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm	What|Is|Data|Mining?|What|Can|Data|Mining|Do|and|Not|Do?| The|Data|Mining|Process|What|Is |Data|Mining?| Data|mining|is|the|practice|of|automatically|searching|large|stores|of|data|to|discover|patterns|and|trends|that|go|beyond|simple|analysis|Data|mining|uses|sophisticated|mathematical|algorithms|to|segment|the|data|and|evaluate|the|probability|of|future|events|Data|mining|is|also|known|as| Knowledge|Discovery|in|Data|(KDD).
4	https://id.wikipedia.org/wiki/Penggalian_data	Penggalian|data|(Bahasa|Inggris|*data*|*mining*)|adalah |ekstraksi|pola|yang|menarik|dari|data|dalam|jumlah|besar|. Suatu|pola|dikatakan|menarik|apabila|pola|tersebut|tidak|sepele,|implisit,|tidak|diketahui|sebelumnya,|dan|berguna.|Pola|yang|disajikan|haruslah|mudah|dipahami,|berlaku|untuk|data|yang|akan|diprediksi|dengan|derajat|kepastian|tertentu,|berguna,|dan|baru|Penggalian|data|memiliki|beberapa|nama|alternative|meskipun|definisi|eksaknya|berbeda,|seperti|KDD|(knowledge|discovery|in|database),|analisis|pola,|arkeologi|data,| pemanenan|informasi,|dan|intelegensia|bisnis.
5	http://www.thearling.com/text/dmwhite/dmwhite.htm	An|Introduction|to|Data|Mining|Discovering|hidden|value|in|your|data|warehouse|Data|mining,|the|extraction|of|hidden |predictive|information|from|large|databases,|is| a|powerful new|technology|with|great|potential|to|help|companies|focus |on |the|most|important|information|in|their|data|warehouses.|Data|mining|tools|predict|future|trends|and|behaviors,|allowing |businesses|to|make|proactive,|knowledge-driven|decisions.|The|automated,|prospective|an

		alyses |offered|by|data|mining|move|beyond|the|analyses|of|past |events|provided|by|retrospective|tools|typical|of|decision|support|systems.|Data|mining|tools|can|answer|business |questions|that|traditionally|were|too|time|consuming|to |resolve.

6. *Filtration of Keyword*
   Setelah teks tersebut dibagi menjadi kata per kata, selanjutnya kata yang mengandung unsur stop word dihapus.

Tabel 6. *Filtration of Keyword*

No	Alamat URL	Konten
1	http://searchsqlserver.techtarget.com/definition/data-mining	Data|mining| |process|sort|through|large|data|set|identify|pattern|establish|relation|solve|problem|through|data|analysis.|Data| mining |tool|allow|enterprise|predict|future|trend.|Download| free|guide|Download|Our|Exclusive|Big|Data|Analytic|Guide|unbiasedlook| real-life|analytic|success|story| include| time|Warner|Cable|case|studytips|evaluate|big|data|tool…
2	https://en.wikipedia.org/wiki/Data_mining	Data| mining|compute|process|discover|pattern|large|data| set|involve|method|intersect|machine|learn|statistic| database|system|interdisciple|field|computer|science |overall|goal| data|mining|process|extract|information|data| |transform|understandable|structure|for |further|use…
3	https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm	Data| mining |Data| mining |Data| mining |Process|Data| mining | Data|mining| |practice| automatically|searching|large|stores| data| discover|pattern|trend|go|beyond|simple|analysis|Data | mining |use|sophisticate|mathematic|algorithm| segment|data|evaluate|probability|future|event|Data | mining |known|Knowledge|Discovery|Data|(KDD)…
4	https://id.wikipedia.org/wiki/Penggalian_data	penggalian|data|Bahasa|Inggris|*data*|*mining*|ekstraksi|pola|data| jumlah| pola|pola| implisit| Pola|disajikan|haruslah| mudah|paham|laku|data|prediksi|derajat|Penggalian| data|nama|alternative| definisi|eksaknya|KDD|knowledge| discovery| database|analisis|pola|arkeologi|data|pemanenan| informasi| intelegensia|bisnis

5	http://www.thearling.com /text/dmwhite/dmwhite.htm	Introduction|Data|Mining|Discovering|hidden|value | data|warehouse|Data|mining|extraction|hidden |predictive|information|large|databases|powerful |technology |great|potential| help|companies|focus |important|information| data|warehouses|Data|mining|tools|predict|future|trends| behaviors|allowing |businesses|make|proactive|knowledge-driven|decisions| automated,|prospective|analyses |offered| data|mining| beyond| analyses| past |events|provided| retrospective|tools|typical| decision|support|systems.|Data|mining|tools| answer|business |questions| traditionally| time|consuming| resolve.

7. *Word Stemming*

Pada tahapan ini berdasarkan kajian teori mengenai *preprocessing word stemming*, peneliti memfasilitasi penggunaan kata Bahasa inggris dan Indonesia pada teks hasil pencarian,stemming mengilangkan previks pada kata hasil pencarian, berikut word stemming untuk *data dari laman* https://en.wikipedia.org/wiki/Data_mining, yang telah melewati proses filtration.

Tabel 7. *Word Stemming*

Kata	Output
Data	*Data*
Mining	*mine*
Data	*Data*
Mining	*mine*
Process	*Process*
practice	*practice*
automatically	*Automatic*
Large	*Large*
Data	*Data*
Sets	*Set*
Involving	*Involve*

Kata	Output
Methods	*method*
Intersection	*Intersect*
Machine	*Machine*
Learning	*Learn*
Statistics	*Statistic*
Database	*Database*
Systems	*System*
interdisciplinary	*Interdicipline*
subfield	*subfield*
computer	*computer*
science	*science*

Pada laman yang berbahasa Indonesia berdasarkan Tabel 1 prefiks dan stop words dihilangkan.

Kata	Output
Penggalian	Gali
Data	Data
Bahasa	Bahasa
Inggris	Inggris
Data	Data
Mining	Mine
Ekstraksi	Ekstrak
pola	Pola
Data	Data
Jumlah	Jumlah
Besar	Besar

Kata	Output
Dipahami	Paham
Berlaku	Laku
Data	Data
Diprediksi	Prediksi
Derajat	Derajat
Penggalian	Gali
Data	Data
nama	Nama
Alternative	Alternative
Definisi	Definisi
Eksaknya	Eksak

Pola	Pola
Dikatakan	Kata
Menarik	Tarik
Pola	Pola
Sepele	Sepele
Implisit	Implisit
Diketahui	Tahu
Berguna	Guna
Pola	Pola
Disajikan	Saji
Haruslah	Harus
Mudah	Mudah

KDD	KDD
Knowledge	Knowledge
Discovery	Discovery
database	database
analisis	analisis
pola	pola
Arkeologi	Arkeologi
Data	Data
Pemanenan	Pemanenan
Informasi	Informasi
Intelegensia	Intelegensia
Bisnis	Bisnis

8. *Indexing*

Pada tahapan ini dilakukan skema pembobotan TF-IDF (*Term Frequency, Inverse Document Frequency*) dengan menghitung beberapa banyak kata kunci muncul dalam laman web tersebut. Dari sample laman yang ditentukan maka koleksi Dokumen |C| = 5, jumlah kemunculan term pada dokumen (TF) , Nt adalah menunjukan jumlah istilah dalam dokumen Nt_i. Tabel pembobotan laman web berdasarkanTF-IDF, dengan total kata setiap dokumen pada simulasi analisis ini adalah 100 kata.

Tabel 8. *Indexing*

Kode dokumen	Url
d1	http://searchsqlserver.techtarget.com/definition/data-mining
d2	https://en.wikipedia.org/wiki/Data_mining
d3	https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm
d4	https://id.wikipedia.org/wiki/Penggalian_data
d5	http://www.thearling.com/text/dmwhite/dmwhite.htm

	tf = jumlah istilah t muncul dlm dokumen / jumlah total kata dlm dokumen					N_t
Q	d1	d2	d3	d4	d5	
data	0.0050	0.0333	0.0750	0.0333	0.0417	0.1883
mining	0.0167	0.0167	0.0583	0.0083	0.0417	0.1417

$\|C\|/N_t$	IDF = LOG(|C| / N_t)	IDF + 1
26.5522	-0.7160	0.2840
35.2941	-0.5315	0.4685

	w = tf(IDF+1)				
Q	d1	d2	d3	d4	d5
data	0.0014	0.0095	0.0213	0.0095	0.0118
mining	0.0078	0.0078	0.0273	0.0039	0.0195
Total Bobot	**0.0092**	**0.0173**	**0.0486**	**0.0134**	**0.0314**

TF Query	IDF + 1	TF*IDF
0.2000	0.2840	0.056799331
1.0000	0.4685	0.468521083

SQRT(wd^2)				
|D1|	|D2|	|D3|	|D4|	|D5|
0.0792	0.2247	0.3371	0.2247	0.2512
0.2348	0.0551	0.1929	0.0276	0.1378
SQRT(wq^2)				
|q1|	|D2|	|D3|	|D4|	|D5|
0.1577	0.4472	0.6708	0.4472	0.5000
0.1414	0.1414	0.2646	0.1000	0.2236

	d1	d2	d3	d4	d5
Dot Product(Q,D) =	0.004	0.005	0.017	0.003	0.012
Cosine Similarity =	0.358	0.05	0.074	**0.028**	0.094

Nilai ambang batas trashhold diambil dari nilai cosine similarity terkecil = ***0.028***

3.4 Text Mining Discovery Menggunakan Algoritma GARW

Seperti pada pemaran kajian pustaka mengenai algoritma apriori, untuk menemukan kombinasi item berdasarkan barang yang dibeli oleh pelanggan , dalam studi kasus ini barang yang dibeli dianalogikan sebagai konten yang dicari pada halaman *web. Association rule* menemukan pola, asosiasi, korelasi yang sering terjadi di antara kumpulan item atau objek dalam database transaksi, database relasional, dan repositori informasi lainnya. Pada kaitan mengenai pencarian di *google*, terdapat motivasi untuk menemukan bagaimana pengguna mendapatkan informasi yang dibutuhkannya saja tanpa melihat keseluruhan konten dalam website. Sehingga hal tersebut terdapat pola yang dipelajari dari umumnya pengguna melakukan pencarian berdasarkan kategori tertentu.

Sebagai gambaran umum, misalkan pada pencarian *google* konten yang dicari dari contoh kasus *web crawler* ini adalah hanya mengenai definisi, metode, latar belakang saja, sehingga hasil pencarian tersebut diharapkan hanya mengenai definisi, metode dan latar belakang saja. *Association rule* akan mengklasifikasikan konten hasil pencarian *google* dengan melakukan penambangan teks hanya yang terkait dengan definisi, metode dan latar belakang saja. Untuk Menganalisis Algoritma GARW yaitu harus :

1. Menentukan nilai ambang batas (bobot *threshold*) seperti pada Tabel pembobotan laman web berdasarkan TF-IDF,
2. ditentukan nilai ambang batas adalah 0.028, artinya laman yang memiliki nilai ambang batas dibawah 0.028 akan di lewati atau tidak diproses,
3. Menyusun Konten URL berdasarkan nilai ambang batas (*bobot threshold*)
4. Generate Association rule menggunakan algoritma Apriori, dan Gunakan minimum *support* dan *confidence*.
5. Dalam proses Generate Association Rule
   a. Scan *list index* dan temukan semua kata kunci yang memenuhi ambang batas *minimum support*,
   b. Dalam K >=2 (K adalah kumpulan kata kunci yang memiliki set k-kata kunci) , kandidat kata kunci C_k ,dimana nilai k di hasilkan dari set kata kunci.

c. Pindai daftar index frekuensi menghitung dari set kata kunci C_k yang dihasilkan pada langkah b,
d. Bandingkan frekuensi set calon kata kunci dengan minimum support,
e. Set kata kunci k yang memenuhi minimum support

Kode dokumen	Konten URL	Item yang di Cari
d3	https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm	Definisi, process
d5	http://www.thearling.com/text/dmwhite/dmwhite.htm	Definisi, *architecture*
d1	http://searchsqlserver.techtarget.com/definition/data-mining	Definisi, *technique*
d2	https://en.wikipedia.org/wiki/Data_mining	Definisi, background, process
d4	https://id.wikipedia.org/wiki/Penggalian_data	Definisi , Latar belakang, Teknik

Iterasi ke-1

Untuk 1-itemset

Itemset	Support Count	Support
Definsi	5	100%
Technique/process	4	80%
Latar belakang / background	2	40%
Arsitektur	1	20%

dapatkan k-itemset dari support yang memenuhi minimum support, kemudian pilih k-itemset sebagai pola frequent tinggi , ditentukan minimum *support* adalah 1.

Itemset	Support Count	Support
Definsi	5	100%
Technique/process	4	80%
Latar belakang / background	2	40%

Iterasi ke-2

Table C2 kombinasi dari k-itemset

Itemset	Support Count	Support
Definsi, Technique/process,	4	80%
Definisi, Latar belakang / background	2	40%
Technique/process, Latar belakang / background	1	20%

Berikut table Pola frequent tinggi diatas minimum support untuk 2-itemset adalah (F2).

Itemset	Support Count	Support
Definsi, Technique/process	4	80%
Definisi, Latar belakang / background	2	40%

Iterasi ke-3

Itemset	Support Count	Support

Definsi, Technique/process, Latar belakang / background	2	40%

Proses berhenti pada iterasi ke 3, pola frequent tinggi yang ditemukan adalah "Definsi, Technique/process, Latar belakang / background"

4. PENUTUP

Association rule menemukan pola, asosiasi, korelasi yang sering terjadi di antara kumpulan item atau objek dalam database transaksi, database relasional, dan repositori informasi lainnya. Penentuan bobot threshold menggunakan *cosine similarity* ini berdasarkan kesamaan istilah antara *query* dengan dokumen konten web (Bathla & Jindal, 2011). Keuntungan dari pendekatan ini adalah bahwa area pencarian menjadi sangat kecil sehingga waktu tunggu pengguna untuk mendapatkan jawaban permintaan berkurang untuk sebagian besar, dan diharapkan dapat mengurangi waktu pencarian informasi dalam memberikan rekomendasi hasil pencarian yang diinginkan berdasarkan pola, asosiasi, korelasi pencarian.

Dengan bobot threshold 0.028 pada simulasi mengeliminasi dokumen web yang berada pada bobot threshold <= 0.028 sehingga proses melakukan *Generate Association Rule* menjadi lebih ringan. Adapun dari sudut pandang konten yang dievaluasi melibatkan multi Bahasa Indonesia dan inggris, hal tersebut menambah variatif konten informasi yang didapatkan. Adapun kekurangan dari algoritma ini adalah hasil *input* pada konten yang akan di evaluasi perlu di lengkapi dengan efisiensi akses data, jika *input* data yang diberikan adalah lebih dari 1000 *record*, maka mekanisme akses perlu di pertimbangkan. Mekanisme akses saat ini adalah dengan memberi limit berdasarkan kelipatan tertentu, Namun penulis perlu menyempurnakan kondisi ini agar dibuatkan lebih dinamis namun efisien sehingga baik dengan jumlah record < 100 atau konten > 1000 *record* memiliki load data yang stabil.

Pada kajian ini foKus pada bagaimana melakukan penambangan teks terhadap dokumen semi terstruktur html dengan mengambil konten yang diperlukan dan kemudian dilakukan penambangan teks untuk menemukan pola. Adapun dokumen yang tidak terstruktur seperti dokumen pdf, doc yang merupakan hasil pencarian menjadi *open issue* untuk penelitian selanjutnya.

DAFTAR PUSTAKA

Bathla, G., & Jindal, R. (2011). Similarity Measures of Research Papers and Patents. *International Journal of Computer Applications*, 10.

Bhujade, V., & Janwe, N. (2011). Knowledge Discovery in Text Mining Technique Using Association Rules Extraction. *International Conference on Computational Intelligence and Communication Systems*, 498-502.

Christopher Olston, M. N. (2010). Web Crawling. *Foundations and Trends in Information Retrieval, 4*(3), 176.

Janwe, V. B. (2011). Knowledge Discovery in Text Mining Technique Using Association Rules Extraction. *International Conference on Computational Intelligence and Communication Systems*, 498-502.

Kulkarni, M. K. (2016). KNowledge Discovery in Text Mining using Association Role Extraction. *International Journal of Computer Application , 143*(12), 31.

Naithani, A. K. (2016). A Comprehensive Study of Text Mining Approach . *International Journal of Computer Science and Networking Security, 16*(2), 69.

Najork, M., & Heydon, A. (2001). *High Performance Web Crawling* . California: Compac System Research Center.

R, J., & Vijayarani, D. S. (2016). Text Mining Reserach: A Survey. *International Journal of Innovative Research in Computer and Communication Engineering*, 6544.

R. Talib, M. K. (2016). Text Mining : Techniques, Applications and Issues. *International Journal of Advanced Computer Science and Applications*, 414.

Sanger, R. F. (2007). *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data.* Cambrige: Cambrige University.

Tala, F. (2003). *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.* Amsterdam: Institute for Logic, Language and Computation The Netherlands.