ISBN: 978-602-52470-4-0
Prosiding Seminar Nasional Multimedia & Artificial Intelligence 2018
Yogyakarta, 6 Oktober 2018

Information Retrival untuk Pencarian Dokumen Tugas Akhir Menggunakan *Sequential Pattern Mining*

Information Retrival for Searching a Final Task Document Using Sequential Pattern Mining

Gusti Ngurah Mega Nata

STMIK STIKOM BALI
Jl. Raya Puputan No.86 Renon - Denpasar, Indonesia
Email: mega@stikom-bali.ac.id

ABSTRAK

Abstrak - Selama ini sistem *information retrieval* menggunakan teknik *text mining* akan menggunakan representasi kata *bag of word*. Pada *Bag of word* setiap kata berdiri sendiri, padahal sebuah *term* bisa terbentuk dari beberapa kata, misal "sistem informasi komputer", "rumah sakit", "sepeda motor", "data mining", *term* tersebut terbentuk dari dua kata atau lebih. *Term* yang terbentuk dari dua kata atau lebih jika menggunakan *bag of word* akan mengilangkan *semantic* dari *term* tersebut. dengan kata lain *bag of word* kurang menjaga semantic dari *term* di dalam dokumen teks. Pada paper ini dilakukan proses *information retrieval* pada dokumen teks dengan memperhatikan urutan dari kata (*sequential of words*) di dalam kalimat. Pembentukan term *sequential of words* akan dilakukan setelah proses *stemming*. Term *sequential of word* yang dibentuk yaitu hanya kata dasar hasil *text preprocessing*. Dokumen teks yang digunakan untuk pengujian yaitu 1000 dokumen skripsi / TA dari mahasiswa. Pada paper ini proses pengalian *sequence of words* pada setiap kalimat yaitu menggunakan *sequential pattern mining*. Hasil dari uji coba yaitu berupa list *sequential of word* yang lebih dari minimum support yang telah ditentukan yaitu 5% dari jumlah kata.

Kata kunci: Information Retrival; Sequential Pattern Mining; Text Mining; tugas akhir;

ABSTRACT

Abstrak-Information retrieval system is using a bag of word representation. In the Bag of Word every word stands alone, while a term can be formed from several words, for example in Indonesia language"sistem informasi komputer" (computer information system), "rumah sakit" (hospital), "sepeda motor" (motorcycle), "data mining", the term is formed of two or more words. Term that is formed from two words or more if using a word bag will remove the semantic from the term. the conclusion is that bag of words does not maintain the semantics of the terms in the text document. In this paper, information retrieval is performed on text documents by observing the order of the words in the sentence. The formation of sequential terms of words will be done after the stemming process. The sequential term of word that is formed is only the basic words of the text preprocessing results. Text documents used for testing are 1000 thesis documents / TA from students. In this paper the process of sequencing the sequence of words in each sentence is using sequential pattern mining. The results of the trial are in the form of a sequential list of words which is more than the minimum support that has been determined which is 5% of the number of words.

Keywords: Information Retrival; Sequential Pattern Mining; Text Mining; tugas akhir;

1. PENDAHULUAN

Menemukan kembali informasi Tugas Akhir (TA) mahasiswa dalam sekumpulan file yang banyak sudah sangat diperlukan di pergutuan tinggi menurut paper (Zuliar Efendi, Mustakim., 2017). Menemukan Tugas Akhir mahasiswa yang sudah lulus sangat membatu team penerimaan usulan tugas Akhir dalam pengecekan *plagiat*, atau mencari rujukan penelitian terkait. Selama ini sistem temu balik menggunakan teknik *text mining* akan menggunakan representasi kata *bag of word* seperti pada paper (Putri Elfa Mas'udia, dkk: 2017). Pada *Bag of word* setiap kata berdiri sendiri menurut buku (Han Jiwai, Kamber, Pei., 2012), Padahal sebuah *term* bisa terbentuk dari beberapa kata misal "sistem informasi komputer", "rumah sakit", "sepeda motor", "data mining", *term* tersebut terbentuk dari dua atau lebih dari dua kata. *Term* yang terbentuk dari dua kata atau lebih jika menggunakan *bag of word* akan mengilangkan semantic dari *term* tersebut. dengan kata lain *bag of word* kurang menjaga semantic dari *term* di dalam dokumen teks.

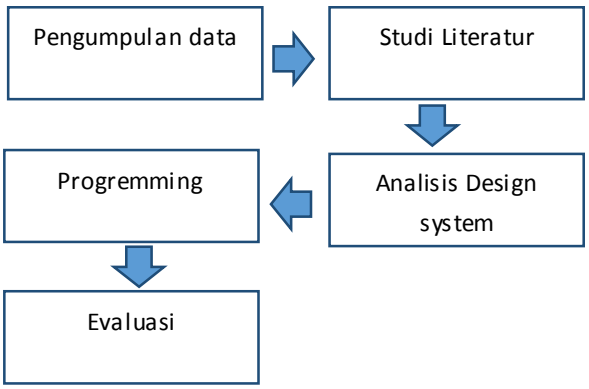
Dalam mencari kembali tugas akhir dalam database pada penelitian ini yaitu dengan cara *information retrieval*. Information retrival dalam penelitian ini memperhitungkan urutan kata (*sequence of words*) yang sering muncul (*frequent*) pada setiap kalimat. *sequence of words* akan di eksplor dengan algoritma *Sequential Pattern Mining* pada fase *feature generation*. Algoritma Sequential pattern mining pada awalnya digunakan untuk mencari hubungan sequential antara satu transaksi dengan transaksi berikutnya yang dilakukan oleh seorang kustomer (Han Jiwai, Kamber, Pei., 2012), namun pada penelitian ini akan diterapkan pada dokumen teks, dimana kalimat akan dijadikan *itemset*, kata dalam kalimat digunakan sebagai *item*, dan satu dokumen berisi sekumpulan *itemsets*.

2. TINJAUAN PUSTAKA

Penelitian dalam bidang *text mining* khususnya dalam proses stemming sudah bayak dilakukan seperti pada paper (Fadillah Z. Tala. 2002) dan (Gede Widnyana putra. 2016). Namun, penelitian yang dilakukan dengan cara menggunakan representasi *term bag of word*. Penelitian yang dilakukan oleh (Widnyana putra. 2016) merupakan penelitian klasifikasi teks menggunakan representasi *bag of word*, dalam proses penelitian tersebut proses *stemming* dilakukan seperti pada *stemming* Bahasa Indonesia menggunakan teknik *forter stemmer*. Penelitian pada paper (Putri Elfa Mas'udia, dkk 2017), yang berjudul : "Information Retrieval Tugas Akhir dan Perhitungan Kemiripan Dokumen Mengacu pada Abstrak Menggunakna Vector Space Model", juga menggunakan representasi *bag of word*. Jadi penelitian yang sudah pernah dilakukan untuk *information retrival* lebih banyak menggunakan potongan kata per kata atau *bag of word* dan masih belum di temukan paper yang menggunakan format *sequential pattern of world* khususnya dalam information retrival pada dokumen tugas akhir berbahasa Indonesia.

3. METODOLOGI PENELITIAN

Metode penelitian yang diterapkan terdiri dari beberapa tahapanan. Berikut adalah tahapan dalam metode penelitian yang digunakan pada penelitian ini:



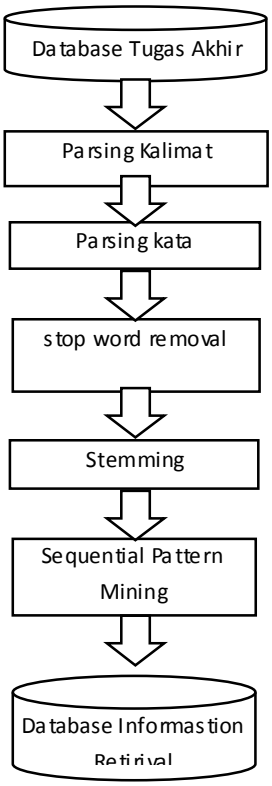
Gambar 1. Alur Penelitian

1. Penggumpulan dokumen tugas akhir di dapat dari tiga prodi. Jumlah dokumen tugas akhir yang digunakna yaitu 1000 dokumen. Namun isi dokumen yang digunakan yaitu abstrak dan latar belakang saja.
2. Studi literature dalam penelitian dan penelusuran penelitian terkait dalam information retrieval dan sequential patter mining.
3. Analisis design system dan preprocessing data teks menjadi refresentasi data keranjang belanja untuk proses sequential pattern mining.

4. Implementasi design system dan pengujian design system terhadap data dokumen teks.
5. Evaluasi hasil pengujian dari design system.

4. PEMBAHASAN

Analisis dari penelitian ini adalah menerapkan *sequential pattern mining* pada dokumen teks untuk mendukung information retrival. Pada proses *preprocessing text* ada perbedaan dengan *preprocessing text mining* yaitu pada penelitian in terdapat parsing kalimat.



Gambar 2. Alur Proses Preprocessing

1. Parsing kalimat

Parsing kalimat yaitu setiap kalimat dalam teks dokumen dipotong menjadi satu *itemset*. Tanda baca yang digunakan untuk memotong kalimat yaitu tanda titik (.), tanda Tanya (?), tanda seru (!), dan baris baru. Perubahan kalimat menjadi itemset untuk proses frequent pattern dan untuk menjadi *sequential pattern*.

Misanya kita memiliki dokumen teks yang diberi kode Doc_01 seperti berikut:

Doc_01:
Data Mining merupakan salah satu teknik penting dalam mencari pengetahuan dalam sekumpulan data digital. Namun, istilah data mining dijadikan kunci utama dalam proses pencarian pengetahuan dalam sekumpulan data oleh para industri, media dan pada lingkungan penelitian. Hal itu karena teknik *data mining* yang memiliki fungsi paling penting yaitu mencari dan menemukan pola menarik yang tersembunyi.

Gambar 3. Contoh Dokumen TA

Pada tabel dibawah ini setiap kalimat dalam Dokumen 01 (Doc_01) sudah diberi kode mulai dari K01,K02…K*n*.

Tabel 1. Hasil Parsing Kalimat

Dkalimat	Konten
K01	Data Mining merupakan salah satu teknik penting dalam mencari pengetahuan dalam sekumpulan data digital.
K02	Namun, istilah data mining dijadikan kunci utama dalam proses pencarian pengetahuan dalam sekumpulan data oleh para industri, media dan pada lingkungan penelitian.
K03	Hal itu karena teknik *data mining* yang memiliki fungsi paling penting yaitu mencari dan menemukan pola menarik yang tersembunyi.

2. Parsing kata

Parsing / *tokenizing* kata baru dilakukan setelah parsing kalimat. Parsing kata memisahkan kata – kata berdasarkan space, koma, symbol, angka dan pemisah lainya. Proses parsing membuat setiap kata menjadi terpisah (*bag of word*) ke dalam list yang disimpan dalam memory dalam bentuk larik. berikut contoh hasil parsing kata dalam kalimat K01:

K01:
Data, Mining, merupakan, salah, satu, teknik, penting, dalam, mencari, pengetahuan, dalam, sekumpulan, data, digital.

Gambar 4. Hasil Parsing kata

Tanda koma pada gambar diatas hanya menandakan bahwa kata – kata tersebut sudah terpisah. Tujuan hasil dari parsing adalah *Bag of word*. *Bag of word* kemudian dilakukan proses *stop word removal*.

3. Stop word removal

Setelah proses parsing / *tokenizing* setiap kata menjadi berdiri sendiri / tidak terikat dengan kata yang lain. Akibat dari pemisahlan kata tersebut, akan ada kata yang tidak memiliki arti yang relevan untuk menentukan ciri dari dokumen yang di *tokenizing*. Kata – kata yang tidak memiliki arti yang relevan tersebut disebut *stop word*. Kumpulan dari stop word disebut *stop list* dan proses untuk menghapus *stop word* dalam dokumen disebut *stopword removal*. dalam kalimat K01 seperti pada gambar 3. Jika kalimat K01 di stopword removal maka akan mengasilkan seperti berikut:

K01:
Data, Mining, merupakan, satu, teknik, penting, mencari, pengetahuan, sekumpulan, data, digital.

Gambar 5. Hasil Stop word removal

4. Stemming

Stemming adalah proses pemetaan dan penguraian berbagai bentuk (*variants*) dari suatu kata menjadi bentuk kata dasarnya. Pada penelitian ini algoritma *stemmer* yang digunakan yaitu algoritma CS. Algoritma ini memerlukan list kata dasar. List kata dasar sebelumnya sudah dimasukkan kedalam database. Jumlah kata dasar yang digunakan yaitu 28526 kata dasar. Berikut adalah hasil stemming:

K01:
Data, Mining, rupa, satu, teknik, penting, cari, tahu, kumpul, data, digital.

Gambar 6. Hasil Stemming

5. Sequential Pattern Mining

Sequence adalah daftar (*list*) terurut dari sekumpulan item (*itemsets*) (Agrawal, dkk. 1995). Jika suatu pola *itemset* sering muncul secara *sequence* pada suatu *dataset* maka disebut *frequent sequential pattern*. Dalam penelitian ini yang dijadikan *itemsets* adalah hasil parsing kalimat, items nya adalah kata dalam kalimat sedangkan sequence yang dicari adalah sequence kata dari masing – masing kalimat. misal I=(i_1, i_2, … i_n) di mana i_j adalah sebuah *item* maka *I adalah itemset*. Sebuah item X dikatakan subset jika, $X \subseteq I$ dimana I adalah *itemsets*. Dan Sebuah *sequence* <a_1 a_2 … a_n> juga dapat terkandung didalam secuence lain <b_1 b_2 … b_m> maka, sequence <a_1 a_2 … a_n> dikatakan subseqence dari <b_1 b_2 … b_m>, atau $a_n \subseteq b_{i_n}$ atau $b_{i_n} \supseteq a_n$ (Agrawal, dkk. 1995) (Ayres Jay. 2002).

Tabel 2. Daftar kalimat

IDdok	ID kalimat	Kata-kata
Doc_01	**K01**	**Data**, **Mining**, rupa, satu, **teknik**, **penting**, **cari**, **tahu**, **kumpul**, **data**, digital.
Doc_01	**K02**	**Data, mining,** kunci, utama, proses, **cari**, **tahu**, **kumpul**, **data,** industri, media, lingkung, teliti.
Doc_01	**K03**	**Teknik**, ***data, mining,*** milik, fungsi, **penting**, **cari**, nemu, pola, narik, sembunyi.

Sebelum proses *sequence of word* setiap kata perlu dicari nilai *frequent pattern* nya. Pada contoh ini nilai minimum frequent pattern yang

Fakultas Teknologi Informasi
Universitas Mercu Buana Yogyakarta
Jl. Jembatan Merah No. 84C, Gejayan, Yogyakarta 55283

diberikan yaitu 2, artinya kata tersebut minimal 2 dua dalam kalimat. Dalam tabel 2 diatas kata yang diblok adalah kata yang memenuhi minimum support.

Setiap *row* pada table 2 adalah sebuah kalimat yang memiliki minimal 1 item / kata. Jadi dapat diartikan bahwa satu kalimat adalah *Itemset*. Dan pada tabel 2 setiap kalimat sudah memiliki kata yang sudah terurut sesuai dengan posisi maka, ini dapat dikatakan satu *sequence*. Maka, sebuah dokumen *D* representasi dari sekumpulan *sequence* (*set of sequence*) dengan kata lain, *D* adalah *sequence representation*. Jika dilihat dari kalimat K01 sequence yang terbentuk yaitu kata **data** memiliki urutan dengan kata – kata di belakangnya, kata mining dengan kata – kata dibelakangnya dan begitu seterusnya. Berkut adalah pembentukkan 2 itemset yaitu:

Tabel 3. Pembentukkan 2-Itemset K01

Item_1	**Item_2**
data	mining
data	teknik
data	penting
data	cari
data	tahu
data	kumpul
data	data
mining	teknik
mining	penting
mining	cari
mining	tahu
mining	kumpul
mining	data
teknik	penting
teknik	cari
teknik	tahu
teknik	kumpul
teknik	data
penting	cari
penting	tahu
penting	kumpul
penting	data
cari	tahu
cari	kumpul
cari	data
tahu	kumpul
tahu	data
kumpul	data

Hasil *sequence* dari *setiap* dokumen pada tabel 2 perlu di batasi untuk mendapatkan *frequent sequential pattern* yang maksimal. Untuk mendapatkan *frequent sequential* pattern maka dilakukan proses seleksi sequence berdasarkan min_support. Dalam contoh kasus ini digunakan >2 dari seluruh kalimat dalam satu dokumen. Berikut adalah sequence pattern yang memenuhi support >2

Tabel 3. Sequence of word dalam kalimat

IDdoc	Kata yang sequence	Jml
Doc01	{data}{mining}	3
Doc01	{data}{cari}	3
Doc01	{mining}{cari}	3

4. KESIMPULAN

Parsing kalimat membantu dalam proses perubahan data teks menjadi itemset. Dengan kalimat menjadi sebuah itemset maka proses *sequential pattern mining* dapat dilakukan pada dokumen teks. *Sequential pattern mining* pada teks berbeda dengan *sequential pattern mining* pada data keranjang belanja, dimana data teks sudah terurut secara item yaitu kata – kata, sedangkan data keranjang belanja belum berurut. Pemanfaatan sequential pattern dalam information retrival dapat menjaga kata – kata yang berurutan dalam kalimat dapat dijaga secara semantic.

5. UCAPAN TERIMA KASIH

Penelitian ini didukung dan dibiayai oleh STMIK STIKOM Bali. Ucapan terima kasih diberikan kepada STMIK STIKOM Bali dan rekan – rekap sesama peneliti atas masukannya.

DAFTAR PUSTAKA

Asian, J., Williams, H. E., Tahaghoghi, S.M.M.,2005, Stemming Indonesian,

Australian Computer Society Inc., Australia.

Even-Zohar, Yair. 2002. Introducing to *Text mining*. Automated Learning Group, University of Illinois.

Fadillah Z. Tala, 2002, "*A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia*", Netherland, Universiteit van Amsterdam,

Gede Widnyana putra, sudarma made, satya kumara. (2016). Klasifikasi Teks Bahasa Bali dengan Metode Supervised Learning Naïve Bayes Classifier. Teknologi Elektro, Vol. 15, No.2.

Han Jiwai, Kamber, Pei., (2012), *Data Mining concepts and techniques third edition*. Morgan Kaufmann publishers

IAN H. Witten, Eibe Frank, Mark A. Hall., (2011), *Data Mining practical machine learning tools and techniques third edition*. Morgan Kaufmann publishers

I wayan simpen. 2008. Afiksasi Bahasa bali: sebuah kajian morfologi generative. SK Akreditasi Nomor: 007/BAN PT/Ak-V/S2/VIII/2006, Vol 15, No.29

Mega Nata Gusti Ngurah, Yudiastra Putu Pande. Preprocessing *Text mining* pada email box berbahasa Indonesia, Konferensi Nasional Sistem & Informatika (KNS&I) 2017

Mega Nata Gusti Ngurah, Yudiastra Putu Pande. Stemming teks sor-singgih Bahasa Bali.

Putri Elfa Mas'udia, dkk (2017), Information Retrieval Tugas Akhir dan Perhitungan Kemiripan Dokumen Mengacu pada Abstrak Menggunakna Vector Space Model. Jurnal SIMETRIS, Vol 8 No 1 April 2017.

Zuliar Efendi, Mustakim., (2017), *Text mining* Classification sebagai rekomentasi dosen pembimbing Tugas Akhir Program Srudi sistem Informasi., Seminar Nasional teknologi Informasi, Kumunikasi dan industri (SNTIKI) 9, Fakultas Sains dan Teknologi, UIN sultan Syarif kasim Riau Pekanbaru, 18 – 19 Mei 2017.