Implementation Method of Maximum Marginal Relevance (MMR) and Steiner Tree Algorithm to Determining Storyline Documents News

Aji Teja Hartanto[1]

[1] Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
Email : itera194@gmail.com[1]

ABSTRACT

In chronological search process a news event online using search engines, users must access the various sites that have relevance for the event. This is because search engines do not provide search results in a structured way. The search process requires conformity result of the query is entered as search criteria. Relevance textual news can be obtained by using a cosine-ringaksan similiarity summary of news by implementing the method of Maximum Marginal Relevance (MMR) which is determined based on the similarity to the query. In the same context, the Search over 90 samples of news documents applied algorithm Steiner Tree in determining the best path (trajectory) in a collection of news documents (vertices) connected as a weighted similarity (side) while the side directed toward a specific document (side directed).

Based on testing usability testing methods directed against six respondents as much as 66.7% of respondents considered that this application tasks more quickly and 75% of users were able to overcome the existing problems with the process of finding a time-saving news. From the performance testing of the algorithms applied, obtained by the complexity of the implementation of the MMR algorithm is O (n2) and Steiner Tree algorithm is O (n). Thus, the implementation of MMR and Steiner Tree in search applications chronological flow of these events can reduce the level of absurdity news anecdotal structured so as to facilitate the readers in determining the understanding of events happening quickly.

Keywords : *text-summarization, MMR, steiner tree, Storyline, Artificial Intellegence (AI).*

1. INTRODUCTION

Based on survey data Markplus Insight and Marketeers Magazine Online 2-13, there are 98% of 2015 respondents use the internet as their main source of information. While the percentage of respondents 58.2% of the content referenced information is news [1]. However, a survey of 41 respondents showed 58.84% of news search takes a long time. This is not independent with search results that do not follow the pattern of the semantics of the amount of data obtained. Thus, the search results are not structured.

The use of MMR commonly used to obtain a representative text (conclusion). This method is used to reduce redundancy in multi perangkingan sentence on the document. In addition, sequencing is structured to follow the news of events in line with the determination of the construction of a network to determine the gas pipelines of a graph. Thus, this study aims to simplify and speed up the search process conducted by readers of online news media in the search for a chronological flow of events moved at maximum marginal relevance application of methods and algorithms steiner tree.

1.1 MMR Method

Algorithm MMR is a summary of the extraction methods used to summarize a single document or multiple documents [2]. MMR document summarizing by calculating the similarity between parts of the text. In peringkasan documents do segmentation process document into sentences and do the grouping in accordance with the sentence gendeer mengkomibasikan MMR is used by a given user query.

1.2 Text Processing

In the process of processing text, there are several stages, including:

1. *Case Folding*, namely the unification process case in a text document.
2. *Tokenizing*, namely the cutting process based on the input string that compose each word [2].
3. *Filtering* & *Stopwords Elimination*, done by taking important words from the token. Do disposal of words that are considered less important *(stop list)*.
4. *Stemming*, is a function to transform the words into a text document into a word dasrnya. In general, basic words in Indonesian consisting of a combination [3]:

Prefix 1 + Prefix 2 + Word base + Suffix 3 + Suffix 2 + Suffix 1

5. *Cosine Similiarity*, the approach used to calculate the relevance of query against the document. Its use is a form of continuation of the TF-IDF weighting process. The equation can be described as follows [4]:

$$COS(\Theta_{ij}) = \frac{\sum_k d_{ik} d_{jk}}{\sqrt{\sum_k d_{ik}^2} \sqrt{\sum_k d_{jk}^2}} \quad (1\text{-}1)$$

6. TF-IDF weighting, *Term Frequency Factor* is the factor that determines the weighting term in a document based on the amount they appear in the document. Meanwhile, IDF (Inverse Document Frequency), namely the reduction of the dominance of terms that often appear in various documents.

   Idf value of a caste system can be calculated using the equation:

$$IDF = log\left(\frac{D}{dfi}\right) \quad (1\text{-}2)$$

   D is the number of documents that contain the term and dfi is the number of occurrences of the term D. As for the algorithm to W (weight) of each document in a keyword that is:

$$W_{d,t} = tf_{d,t} * IDF_t \quad (1\text{-}3)$$

Users who want a sample space of information around the query, then it should set λ at a lower value. As for users who want to focus on strengthening the documents more relevant, it should be set at a value closer to λ. Sentences with the highest MMR value of each calculation iteration will be taken, then chosen as a summary. Iteration stops when the maximum MMR results equal to 0 [2].

1.3 Graph Theory

Graf is defined as a system consisting of two components, ie the set of nonempty V (G) whose members are called point and set the E (G) in the form set unsorted couple of pieces of different points in V (G) [5]. There are two general types of graphs:

1. *Undirected Graph*, is the sides do not have a sense of direction
2. *Directed Graph*, is the sides have a sense of direction.

The bipartite graph is a graph whose vertices can be divided into two sets, for example, part of U and V, so that each side of the graph that connects a node in U with a knot on V, the graph can be expressed as G (U, V). so even with the knot on V [6].

1.4 Multiview Graph

Multi-view graph merupakan G = (V,E,A) yakni graf yang terdiri dari *verteks*, *edge* tidak berarah dan seperangkat *edge* terarah *(arcs).*

1.5 Dominating Set

Dominating Set (Association of Domination) is a concept in the determination of a point on a graph with the provisions Dominating set point as the cover or cover a point that is around and a minimum with the provisions simple graph that does not have a loop and double sided.

1.6 Steiner Tree Algorithm

Steiner tree algorithm is a set of subsets or certain parts of vertices spanning tree (spanning tree).

1.7 Steiner Tree Algorithm

A chronological flow is another name for advanced groove, straight groove or a groove progressive. The events displayed in chronological order, go ahead, coherently from the beginning of the middle to late stages. In this groove are a matter of hours, minutes, seconds, day and so forth.

2. Contents of Research

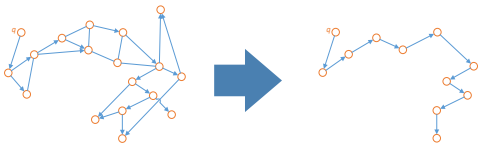
2.1 Analysis of System

Following is a description of a description of the system to be built::

1. In the first stage there are several documents in the database news.
2. Users will provide feedback or query to start the process.
3. Then do the summarization of each document in the news for taking core phrase that represents the document. At this stage method is used maximum marginal relevance. Summarization process is calculated with a given query by the user.
4. After getting a sentence that represents the document among the documents relevant news then performed attribution in accordance with the topic structurally. Each document is represented by a representative sentence node or vertices (points / vertices)

with containing weight. When the weight is high then they have a high similarity with the query.

5. The next step is the formation of the multi-view graph graph formed both with the side (edge) trending and non-trending.
6. The next set specified domination that will cover (covering) of neighboring nodes which become part of the set minimum dominance by using a bipartite graph
7. The next step is to connect the dots that are on the set of such dominance that it becomes a steiner tree.



Picture 2.1 Steiner tree forming process [7]

8. Once formed Steiner Tree, the data presented in tabular form in sequence according timeline of events when the news is rumored.

2.2 Analisis MMR

Given a word document to analyze the MMR algorithm to determine the summary document as follows:

Washington (ANTARA News) - Presiden AS George W. Bush akan mengirim hingga 40.000 tentara lagi ke Irak saat ia mengumumkan revisi kebijakan soal Irak, tulis media massa AS, Kamis. Laporan-laporan media memberikan perkiraan antara 9.000 dan 40.000 tambahan tentara akan dikirim ke Irak. Saat ini menurut sumber militer ada 130.000 tentara AS di Irak. Rencana tersebut mungkin suatu kontroversial karena kebijakan perang Irak makin tidak populer di mata publik AS. Jaringan televisi CNN melaporkan, Bush bakal mengirim 20.000 hingga 40.000 tentara lagi. Pengumumannya kemungkinan pada awal pekan depan. Suatu peningkatan kekuatan pasukan merupakan hal yang secara aktif dibicarakan, kata seorang pejabat senior di pemerintahan kepada CNN. Sementara itu CBS News yang mengutip sumber militer AS, mengatakan bahwa Bush tengah mempersiapkan pengiriman sekitar 9.000 tentara dan marinir ke Irak, sementara 11.000 lainnya bersiaga di Kuwati dan AS. Dua brigade tentara dengan jumlah sekitar 7.500 personil akan berangkat ke Baghdad, sedangkan sekitar 1.500 marinir ke provinsi Al-Anba, lapor CBS. Pasukan lainnya bersiaga di Kuwait, dan dua lagi di AS. Surat kabar The McClatchy melaporkan bahwa Bush sedang mempertimbangkan pengiriman tiga hingga empat brigade tempur AS, atau antara 15.000 dan 20.000

Assuming the query entered by the user is *"Bagaimana terjadinya peperangan di suriah dan irak?"*. From these documents, obtained by the segmentation of the sentence as follows:

Table 2.1 Segmentation of Sentence

No	Kalimat
1	Washington (ANTARA News) - Presiden AS George W
2	Bush akan mengirim hingga 40.000 tentara lagi ke Irak saat ia mengumumkan revisi kebijakan soal Irak, tulis media massa AS, Kamis
3	Laporan-laporan media memberikan perkiraan antara 9.000 dan 40.000 tambahan tentara akan dikirim ke Irak
4	Saat ini menurut sumber militer ada 130.000 tentara AS di Irak
5	Rencana tersebut mungkin suatu kontroversial karena kebijakan perang Irak makin tidak populer di mata publik AS.
6	Jaringan televisi CNN melaporkan, Bush bakal mengirim 20.000 hingga 40.000 tentara lagi
7	Pengumumannya kemungkinan pada awal pekan depan
8	Suatu peningkatan kekuatan pasukan merupakan hal yang secara aktif dibicarakan, kata seorang pejabat senior di pemerintahan kepada CNN
9	Sementara itu CBS News yang mengutip sumber militer AS, mengatakan bahwa Bush tengah mempersiapkan pengiriman sekitar 9.000 tentara dan marinir ke Irak, sementara 11.000 lainnya bersiaga di Kuwati dan AS
10	Dua brigade tentara dengan jumlah sekitar 7.500 personil akan berangkat ke Baghdad, sedangkan sekitar 1.500 marinir ke provinsi Al-Anba, lapor CBS
11	Pasukan lainnya bersiaga di Kuwait, dan dua lagi di AS
12	Surat kabar The McClatchy melaporkan bahwa Bush sedang mempertimbangkan pengiriman tiga hingga empat brigade tempur AS, atau antara 15.000 dan 20.000 tentara, demikian AFP

Phase text processing is done to get the root of the wording contained in the article. Here are the stages:

1. Case Folding, at this stage, the entire sentence is transformed into lowercase without capital letters.
2. *Tokenizing*, 8 segments taken as an example of the results of tokenizing.

Table 2.2 *Tokenizing*

No	Segmen 8
1	suatu
2	peningkatan
3	kekuatan
4	pasukan
5	merupakan
6	hal
7	yang
8	secara
9	aktif
10	dibicarakan
11	kata
12	seorang
13	pejabat
14	senior
15	di
16	pemerintahan
17	kepada
18	cnn

3. *Filtering* & *Stopwords Elimination*, words that are not too important to be eliminated or with other terms elimination of stopwords.
4. *Stemming*, ECS used at this stage to stage stemmer Stemmer in which change every string in the form of basic words. Then, is obtained:

Table 2.3 *Stemming*

No	Contoh Kalimat Segmen 8
1	Tingkat
2	Kuat
3	Pasukan
4	aktif
5	bicara
6	kata
7	jabat
8	senior
9	pemerintah
10	Cnn

Thus, from the stage of text processing, text processing results obtained from:

Table 2.4 Results of Text Processing

No	Segmen 1	Segmen 2	Segmen 3	Segmen 4	Segmen 5	Segmen 6
1	*Washington*	*Bush*	*lapor*	*sumber*	*rencana*	*Jaring*
2	*antaranews*	*Kirim*	*media*	*militer*	*kontroversial*	*televisi*
3	*Presiden*	*40000*	*kira*	*130000*	*bijak*	*Cnn*
4	*As*	*tentara*	*9000*	*tentara*	*perang*	*Lapor*
5	*George*	*Irak*	*40000*	*as*	*irak*	*Bush*
6	*W*	*Umum*	*tambah*	*irak*	*populer*	*Kirim*
7		*Revisi*	*tentara*		*mata*	*20000*
8		*Bijak*	*kirim*		*publik*	*40000*
9		*Irak*	*irak*		*as*	*tentara*
10		*Tulis*				
11		*Media*				
12		*Massa*				
13		*As*				
14		*Kamis*				

No	Segmen 7	Segmen 8	Segmen 9	Segmen 10	Segmen 11	Segmen 12
1	*Umum*	*Tingkat*	*cbs*	*Brigade*	*pasukan*	*Surat*
2	*Pekan*	*Kuat*	*news*	*Tentara*	*siaga*	*Kabar*
3		*Pasukan*	*kutip*	*7500*	*kuwait*	*The*
4		*Aktif*	*sumber*	*personil*	*as*	*mcclatchy*
5		*Bicara*	*militer*	*berangkat*		*Lapor*
6		*Kata*	*As*	*baghdad*		*Bush*
7		*Jabat*	*kata*	*1500*		*timbang*
8		*Senior*	*bush*	*Marinir*		*Kirim*
9		*Pemerintah*	*siap*	*provinsi*		*brigade*
10		*Cnn*	*kirim*	*al-anba*		*tempur*
11			*9000*	*Lapor*		*As*
12			*tentara*	*Cbs*		*15000*
13			*marinir*			*20000*
14			*irak*			
15			*11000*			
16			*siaga*			
17			*Kuwait*			
18			*As*			

Table 2.4 Count TF dan IDF

Term	Q	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	SF	LOG(n/sf)
As		1	1		1	1				2		1	1	$\sum TF = 8$	Log(13/8)=0.211

Setiap jumlah kemunculan term pada masing-masing segmen dikalikan dengan nilai *IDF*. Maka diperoleh bobot term sebagai berikut:

Table 2.5 P Term Weight calculations "US"

	Q	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
TF	0	1	1	0	1	1	0	0	0	2	0	1	1
IDF	0.211	0.211	0.211	0.211	0.211	0.211	0.211	0.211	0.211	0.211	0.211	0.211	0.211
W(TF) = TF*IDF	0	0.211	0.211	0	0.211	0.211				0.422		0.211	0.211

After that, the overall weighted summation term on a particular segment, the S1 is obtained as follows:

$\sum$ W(TF)	5.781

Thus, the weight of the segment as a whole, can be seen as follows:

Table 2.6 The overall result of the weight of the segment.

Q	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
3.377	5.781	9.502	6.09	3.924	7.743	6.266	1.751	10.237	12.981	11.085	2.65	11.074

Furthermore, the calculation of scalar multiplication between the query with each document (sentence segmentation) on the S2 segment:

$$\sum(W(Q) * W(S2)) = 0.226$$

The second phase calculates the length of each document including a query by the square of the weighted term of each document, do the sum of the squared values and then rooted to produce the

denominator in the equation 2.1. For example take the value segment of the sentence S1:

$\sqrt{\sum W(TF)^2}$ pada segmen 2 adalah $\sqrt{6.25}$ = 2.5

Thus, the overall result, as follows:

Table 2.7 Calculation " vector lengths "

Q	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
1.804	2.5	2.853	2.207	1.768	2.772	2.24	1.283	3.266	3.33	3.323	1.424	3.279

Then, apply the formula for calculating the Cosine Similarity similarity query with the segmentation of other sentences. In S2, as follows:

$$\text{COS (Q, S2)} = \frac{0.672}{1.804 * 2.853} = 0.044$$

Thus, Retrieved:

Table 2.8 Similiarity Value Between Segments

	Q	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
Q	**1**	**0**	**0.044**	**0.028**	**0**	**0.155**	**0**	**0**	**0**	**0.019**	**0**	**0**	**0**
S1	**0**	1	0.006	0	0.01	0.006	0	0	0	0.011	0	0.013	0.005
S2	**0.044**	0.006	1	0.251	0.112	0.118	0.149	0.111	0	0.091	0.012	0.011	0.051
S3	**0.028**	0	0.251	1	0.029	0.018	0.193	0	0	0.144	0.051	0	0.06
S4	**0**	0.01	0.112	0.029	1	0.009	0.029	0.179	0	0.259	0.019	0.018	0.008
S5	**0.155**	0.006	0.118	0.018	0.009	1	0	0	0	0.022	0	0.011	0.005
S6	**0**	0	0.149	0.193	0.029	0	1	0	0.09	0.073	0.05	0	0.185
S7	**0**	0	0.111	0	0.179	0	0	1	0	0	0	0	0
S8	**0**	0	0	0	0	0	0.09	0	1	0.061	0	0.142	0
S9	**0.019**	0.011	0.091	0.144	0.259	0.022	0.073	0	0.061	1	0.13	0.298	0.048
S10	**0**	0	0.012	0.051	0.019	0	0.05	0	0	0.13	1	0	0.085
S11	**0**	0.013	0.011	0	0.018	0.011	0	0	0.142	0.298	0	1	0.01
S12	**0**	0.005	0.051	0.06	0.008	0.005	0.185	0	0	0.048	0.085	0.01	1

MMR calculation method is done by counting the iterations between any combination of two matrix cosine similiarity query the relevance and similiarity sentence. When the sample space information will be taken around the query, then it should set λ at a lower value. However, if you want to focus on strengthening the relevant documents then set λ at a value closer to λ. Sentences with the value of the highest MMR sentence that will be representative of textual news articles.

$$\boldsymbol{MMR} = \boldsymbol{arg} \max_{\boldsymbol{D_1 \epsilon R \backslash S}} \left[\boldsymbol{\lambda\, Sim_1(D_i, Q)} - \boldsymbol{(1-\lambda)} \max_{\boldsymbol{D_j \epsilon S}} \boldsymbol{Sim_2\,(D_i, D_j)} \right] \quad (2\text{-}1)$$

From the analysis, the results obtained iteration as follows:

Table 1.1 Iteration

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
Q1	0	0.044	0.028	0	0.155	0	0	0	0.019	0	0	0
S1	0		0	-0.002	-0.0012							
S2	0.0352		-0.015		0.0116							
S3	0.0224				0.0188							
S4	0		-0.0058		-0.0018							
S5	0.124											
S6	0		-0.0386		0							
S7	0		0		0							
S8	0		0		0							
S9	0.0152		-0.0136		0.0108							
S10	0		-0.0102		0							
S11	0		0		-0.0022							
S12	0		-0.012		-0.001							

Table 2.9 Results iteration MMR

Iterasi ke	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
1	0	0.035	0.022	0	**0.124**	0	0	0	0.015	0	0	0
2	-	0.012	**0.0188**	-		0	0	0	0.0108	0	-	-
3		-		-		-	0	0	-	-	0	-

Table 2.10 Thickness Maximum MMR in MMR iteration.

Iterasi ke	**Kalimat**	**Bobot ArgMax *MMR***
***MMR*MAX(1)**	S5	0.124
***MMR*MAX(2)**	S3	0.0188

From the calculation of MMR, then obtained by the extraction of the following sentences:

Table 2.11 MMR Extraction Results

Kalimat ke	**Kalimat**
S5	rencana tersebut mungkin suatu kontroversial karena kebijakan perang irak makin tidak populer di mata publik as.
S3	laporan-laporan media memberikan perkiraan antara 9.000 dan 40.000 tambahan tentara akan dikirim ke irak

2.3 Steiner Tree Algorithm Analysis

As a continuation of the MMR calculations, given the case that there are several documents that news is news documents in the same topic that is of international events. It is assumed that the results obtained from peringkasan is the same query to the previous case. This news gathered from news portals as already mentioned sebelumyan. Collection of news can be presented as follows:

Table 2.12 Document *Sumamry*

No.	***Timestamp***	**Konten Berita Tekstual**
1	4 Januari 2007 22:43 WIB	*laporan-laporan media memberikan perkiraan antara 9.000 dan 40.000 tambahan tentara akan dikirim ke irak*
2	14 Des 2003; 23:14 WIB	*Presiden Irak Saddam Hussein ditangkap dalam sebuah operasi bersandi "Red Dawn" atau Fajar Merah*

3	13 Nov 2002 02:18 WIB	*Presiden Amerika Serikat George Walker Bush mengancam akan mengerahkan seluruh kekuatan militer AS untuk menggempur Irak.*
4	20 Maret 2003; 19:35 WIB	*Pasukan Amerika Serikat dan sekutunya telah memulai perang dengan menggempur daerah selatan Kota Baghdad, Irak, Kamis (20/3) pagi waktu setempat*
5	18 Juli 2003; 19:57 WIB	*Peringatan 35 tahun kekuasaan Partai Baath di Kota Fallujah, Irak, berlanjut.*
6	on 23 Nov 2004 06:59 WIB	*Komisi Pemilihan Umum Ukraina menyatakan Perdana Menteri Viktor Yanukovych sebagai pemenang pemilihan umum. Yanukovych yang didukung pemerintah Rusia meraup suara 49,42 persen.*
7	on 09 Des 2000 08:07 WIB	*Dua ledakan bom yang terjadi hampir bersamaan mengguncang sebuah kota di perbatasan Rusia-Chechnya, Jumat (8/12)*
8	on 05 Okt 2001 at 19:00 WIB	*Sebuah pesawat jenis Tupolev 154 milik Maskapai Penerbangan Rusia Sibir Airlines meledak dan jatuh di atas Laut Hitam, Kamis (4/10).*
9	14-4-2003 / 20:52 WIB	Tentara Amerika Serikat tiba di Mosul sejak Jumat dengan 11 mobil, diikuti oleh sekitar 300 tentara Kurdi.
10	17 Oktober 2007 14:20	*Pemerintah wilayah otonom Kurdi di Irak Utara memperingatkan para anggota parlemen Turki bahwa campur tangan apapun adalah ilegal.*
11	5 April 2006 09:46	*Pemerintah pusat Baghdad, yang menyebut operasi militer itu sebagai upaya meredam pemberontakan gerilyawan Kurdi, menyatakan wilayah itu sebagai kawasan terlarang.*
12	22 Oktober 2015 20:09	*Assad bertemu dengan Vladimir Putin di Moscow Selasa waktu setempat dan berterimakasih kepada Putin karena telah melancarkan serangan udara kepada lawan-lawannya di Suriah.*
13	25 Oktober 2015, 13:34 WIB	*AS diketahui memang telah menolak proposal yang diajukan Rusia untuk menyatukan kekuatan melawan ISIS.*
14	15-4-2003 / 17:11 WIB	*Saat berpidato di Gedung Putih, Presiden AS George W. Bush mengatakan, Suriah memiliki senjata pemusnah massal.*
15	30/09/2015 18:31 WIB	*Parlemen Rusia akhirnya membulatkan suara, memberikan kewenangan bagi pemerintah Presiden Vladimir Putin untuk mengerahkan kekuatan militer negara ke Suriah.*
16	Sabtu, 04 Peb 2012 08:02:16 WIB	*Sekitar 130.000 prajurit NATO bekerja sama dengan lebih dari 300.000 anggota pasukan keamanan Afghanistan.*

By following the pattern of processing cosine similarity, then gained weight similarity between documents as follows:

Table 2.13. Weight of similarity between the document summary

Q	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16
1	0.026	0.024	0.028	0.163	0.026	0	0	0	0	0.022	0	0.057	0	0.061	0.069	0
0.026	1	0.012	0.014	0.012	0.013	0	0	0	0.096	0.011	0	0	0	0	0	0
0.024	0.012	1	0.047	0.011	0.012	0	0	0	0	0.01	0.058	0	0	0.028	0.032	0
0.028	0.014	0.047	1	0.188	0.014	0	0	0	0.103	0.012	0.045	0	0.115	0.225	0.228	0
0.163	0.012	0.011	0.188	1	0.012	0	0	0.054	0.089	0.01	0.059	0	0	0	0	0.085
0.026	0.013	0.012	0.014	0.012	1	0	0.081	0	0	0.078	0	0	0	0	0	0
0	0	0	0	0	0	1	0.017	0.013	0	0.02	0.02	0	0.018	0	0.092	0
0	0	0	0	0	0.081	0.017	1	0.08	0.076	0	0	0	0.028	0	0.026	0
0	0	0	0	0.054	0	0.013	0.08	1	0	0	0	0	0.02	0.051	0.019	0
0	0.096	0	0.103	0.089	0	0	0.076	0	1	0.041	0.04	0	0	0	0	0
0.022	0.011	0.01	0.012	0.01	0.078	0.02	0	0	0.041	1	0.116	0	0	0	0.097	0.079
0	0	0.058	0.045	0.059	0	0.02	0	0	0.04	0.116	1	0	0	0	0.071	0
0.057	0	0	0	0	0	0	0	0	0	0	0	1	0.068	0.025	0.176	0
0	0	0	0.115	0	0	0.018	0.028	0.02	0	0	0	0.068	1	0.047	0.081	0
0.061	0	0.028	0.225	0	0	0	0	0.051	0	0	0	0.025	0.047	1	0.062	0
0.069	0	0.032	0.228	0	0	0.092	0.026	0.019	0	0.097	0.071	0.176	0.081	0.062	1	0
0	0	0	0	0.085	0	0	0	0	0	0.079	0	0	0	0	0	1

The weight of the non-directional edge is the value of the similarity of the two edge interrelated. The weight of each edge can be seen in the form of adjcency matrix.

Table 2.14 Adjacency Matrix from *Undirected Graph*

Q	R1	R2	R3	R4	R5	R6	R7	R8	R9	R1	R11	R12	R13	R14	R15	R16
1	1	1	1	1	1					1		1		1	1	
1	1	1	1	1	1				1	1						
1	1	1	1	1	1					1	1			1	1	
1	1	1	1	1	1				1	1	1		1	1	1	
1	1	1	1	1	1			1	1	1	1					1
1	1	1	1	1	1		1			1						
						1	1	1		1	1		1		1	
					1	1	1	1	1				1		1	
				1		1	1	1					1	1	1	
	1		1	1			1		1	1	1					
1	1	1	1	1	1	1			1	1	1				1	1
		1	1	1		1			1	1	1				1	
1												1	1	1	.1	
			1			1	1	1				1	1	1	1	
1		1	1					1				1	1	1	1	
1		1	1			1	1	1		1	1	1	1	1	1	
				1						1						1

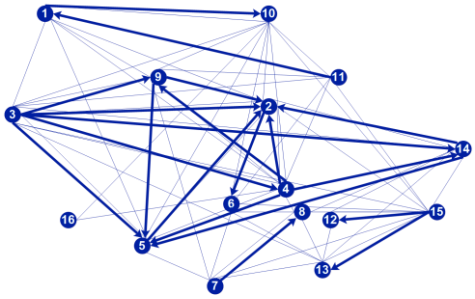
Based on the specified grace period and the value of the cumulative index of the incident, obtained a directed graph of the similarity table in the form adjacency matrix, as follows:

Table 2.15 Adjacency Matriks from *Directed Graph*

	Q	R1	R2	R3	R4	R5	R6	R7	R8	R9	R1	R11	R12	R13	R14	R15	R16
Q																	
R1											1						
R2							1										
R3			1		1	1									1		
R4			1			1				1							
R5			1														
R6																	
R7								1									

R8															
R9			1			1									
R10															
R11		1													
R12															
R13															
R14			1			1									
R15												1	1		
R16															

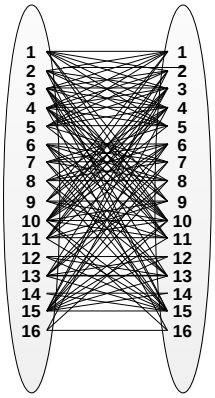
From the table above can be in Picturekan multi-view graph as follows:



Picture 2.2 *Multiview Graph Representation*

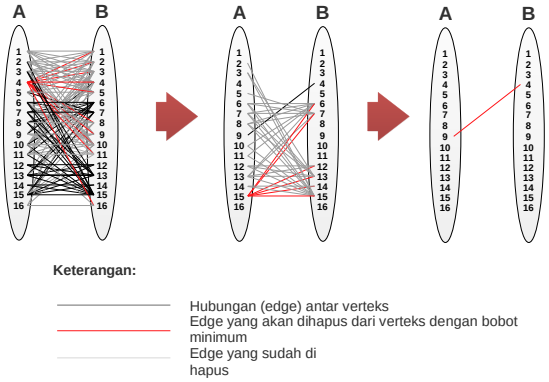
2.4. Deterimining Minimum Dominating Set

In determining the set of domination, first change the graph above into a bipartite graph is assumed as a graph R. This is done to calculate covering numbers (number of points / vertex cover).



Picture 2.3. Bipartite Graph Form

did return the next stage as in the early stages of the search for the vertex with the greatest weight on a set A then remove the edge associated with the selected vertex. It can be seen the following results on graph R:

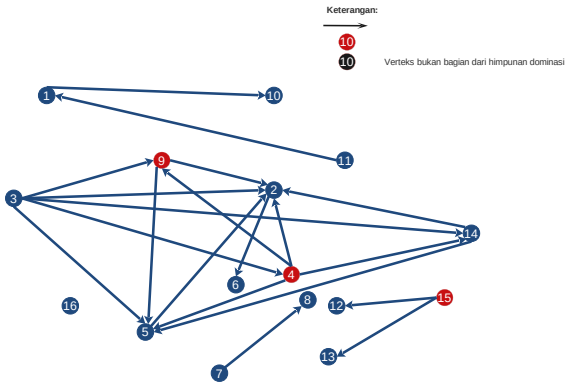


Picture 2.4. Finding Dominating Set

DS (*Dominating Set*) = {R4, R15, R9}

2.5. Determining Steiner Tree

Having obtained the dominated points, we then performed approximation steiner tree which is to connect the points (vertices), which becomes part of the set of domination. Here is a picture graph with edge defined by a set of vertices and their timestamps weighted domination.



Picture 2.5. Directed Graph

To form a steiner tree, weighted matrix and the matrix takes the shortest path.

Here is a table of similarity based query:

Table 2.16 Table Similarity Query

	R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**	**R11**	**R12**	**R13**	**R14**	**R15**	**R16**
Q	0.026	0.024	0.028	0.163	0.026	0	0	0	0	0.022	0	0.057	0	0.061	0.1	0
1-w(v)	0.974	0.976	0.972	0.837	0.974					0.978		0.943		0.939	0.931	

Of the weight gain which is calculated based on the query. Thus, the value 1-W (v) is the weight given to each vertex representing the weight of a directed edge (A). Thus, the results can be seen as follows:

Table 2.17 Table Tracks (A) with Weight of Vertex

No	v	**R7**	**R8**	**R3**	**R4**	**R9**	**R14**	**R5**	**R2**	**R6**	**R11**	**R1**	**R10**	**R16**	**R15**	**R12**	**R13**
1	**R7**		**0.92**														
2	**R8**																
3	**R3**				**0.812**		**0.775**	**0.986**	**0.953**								

4	**R4**				**0.911**		**0.988**	**0.989**								
5	**R9**						1	1								
6	**R14**						1	**0.972**								
7	**R5**							**0.988**								
8	**R2**								1							
9	**R6**															
10	**R11**										1					
11	**R1**											**0.989**				
12	**R10**															
13	**R16**															
14	**R15**														**0.824**	**0.919**
15	**R12**															
16	**R13**															

Here is the application of Steiner Tree algorithm:

Steiner Tree Algorithm

Input:

a. V adalah vertex, W is total of Weight and A is total of Arc *(directed edge)*

b. S is set of terminal (dominated vertex), q is vertex of query and k is position value of v which is part of S or not.

c. UG is *Undirected Graph*

d. DG is *Directed Graph*

Output: Steiner tree T, the roots of the q covering at least k vertices in S

$A_i(k, v_0, UG, DG, S)$
T ← ϕ (kosong)
harga(T) ← 0
if k < **|S|** do
  Tbest ← ϕ (kosong)
  harga(Tbest) ← 0
  T' ← ϕ (kosong)
  harga(T') ← 0
  harga ← ∞
  for v ∈ UG ≠ 0, (v_0, v) ∈ DG, 0 ≤ k' < **|V|** do
    T' ← $A_i(k, v, UG, DG, S)$ ∪ {(v0, v)}
    if harga > cost(T') then
      Tbest ← T'
  T ← T ∪ Tbest
return T

Thus, the results obtained construction Steiner Tree:



Picture 2.6. *Storyline (Steiner Tree)*

Graf Steiner Tree gained interpreted as a chronological flow of a collection of vertices with a directed edge based on the time of the incident. Vertex selected as part of the graph steiner tree is an accumulation of vertex weights with weights highest similarity to the query. Thus, a chronological flow (Storyline) obtained are as follows:

Table 2.18 *Storyline*

V	*Timestamp*	*Ringkasan Dokumen Query Focussed*	*Bobot W(v)*
R4	*20 Maret 2003; 19:35 WIB*	*Pasukan Amerika Serikat dan sekutunya telah memulai perang dengan menggempur daerah selatan Kota Baghdad, Irak, Kamis (20/3) pagi waktu setempat*	*0.163*
R9	*14-4-2003 / 20:52 WIB*	*Tentara Amerika Serikat tiba di Mosul sejak Jumat dengan 11 mobil, diikuti oleh sekitar 300 tentara Kurdi.*	

2.5. Testing

In tests conducted using the method of usability testing, measurement determined based on the recommendations of the respondents in the usability aspect. From the level of learnability, gained as much as 66.65% provides convenience in completing tasks. Based on Efficiency aspect, 66.7% of respondents considered that this application tasks more quickly. From the memorable, as much as 100% of respondents feel the ease in accessing information. In the aspect of error, the user is able to overcome the problems that exist with news search process that saves time with a backlog of as much as 75%. While aspects of satisfication, as much as 83.33% are satisfied with this application because according to user needs.

From these data acquisition, then the accumulation of usability scale levels obtained percentage of total as much as 75%, which is based on the quantitative tables in the category of successful qualification, meaning that the development of this system has been successful.

Moreover, performance testing Pursuant to the algorithm applied, obtained by the complexity of the implementation of the MMR algorithm is O (n2) and Steiner Tree algorithm is O (n).

3. CONCLUSION

Based on the results of the implementation and testing of methods and algorithms MMR Steiner Tree, the obtained conclusion that search chronological flow efficiently provide convenience to the user in searching for news sequentially in the form of a chronological flow.

4. BIBLIOGRAPHY

[1] E. Lukman, "Laporan: Inilah Yang Dilakukan 74,6 Juta Pengguna Internet Indonesia Ketika Online," Techinasia, 31 Oktober 2013. [Online]. Available: Https://Id.Techinasia.Com/Tingkah-Laku-Pengguna-Internet-Indonesia/. [Accessed 7 10 2015].

[2] M. Mustaqhfiri, Z. Abidin And R. Kusumawati, "Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance," *Matics*, Vol. Iv, No. 4, Pp. 1-14, 2011.

[3] D. Nopiyanti And K. A. Sekarwati, "Aplikasi Pencarian Kata Dasar Dokumen Berbahasa Indonesia Dengan Metode Stemming Porter Menggunakan Php & Mysql," In Seminar Ilmiah Nasional Komputer Dan Sistem Intelijen (Kommit 2014) , Depok, 2014.

[4] R. V. Imbar, A. M. Ayub And A. Rehatta, "Implementasi Cosine Similarity Dan Algoritma Smith-Waterman Untuk Mendeteksi Kemiripan Teks," Informatika, Vol. 10, No. 1, Pp. 1-12, 2014.

[5] R. Nicholas, "Aplikasi Graf Berbobot Dalam Menentukan Jalur Angkot (Angkutan Kota) Tercepat," Itb, Bandung, 2010/2011.

[6] Z. Fathoni, "Algoritma Penentuan Graf Bipartit," Makalah If2091, P. 2, 2009.

[7] C. Lin, C. Lin And J. Li, "Generating Event Storylines From Microblogs," In Cikm '12 Proceedings Of The 21st Acm International Conference On Information And Knowledge Management, New York, 2012 .