Indonesian Automated Text Summarization

Gregorius S. Budhi[1], Rolly Intan[2], Silvia R[3]., Stevanus R. R.
Petra Christian University, Informatics Engineering Dept.
Siwalankerto Street 121 - 131, Surabaya 60236
e-mail: greg@petra.ac.id[1], rintan@petra.ac.id[2], silvia@petra.ac.id[3]

Abstract

Automated Text Summarization (ATS) is a computer-based application to produce a summary from an article but it still keeps an accurate main point from the content of the article. In this research, we build an Indonesian ATS aplication using a virtual graph concept, calculation weight of sentence and weight of relation among sentences, Deductive – Inductive method in Indonesian Language and Exhaustive Shortest Path Algorithm to provide a sumarization path from the first sentence to the last sentence on every paragraph in the article. The test result shows that the quality of summarization result depends on the type and the structure of the article. System will produce a good summary if the type of the article is an argumentation type and the article's structure has many paragraphs in which each paragraph has more than two sentences.
Keywords: *Automated Text Summarization, Indonesian Deductive-Inductive paragraph, Dijkstra's Algorithm*

1. Introduction

Automated Text Summarization (ATS), simply called Text Summarization, is a computer-based application that summarizes an article by which the results of the applications still keep the main point of the article. The main goal is to show only some main sentences of the article as a summary, and hopefully by reading the sentences, readers will save their time in understanding as much as possible the whole content of the article.

In this research, we implement text summarization using type of *informative summaries* that only concerns to provide information based on the important aspects of the article by searching and including as much as possible all relevant topics. The summaries will ignore the irrelevant topics and unimportant detail or supporting information. They only produce the important section from the article. Generally, the type of summaries is used for the overview of article.

2. Basic Theory

2.1. Graph

A graph G = (V, E) consists of a set of vertices, V, and a set of edges, E. Each edge is a pair (v, w), where v, w ∈ V. Edges are sometimes referred to as arcs. If the pair is ordered, then the graph is called directed graph or *digraph*. Vertex w is adjacent to v if and only if (v, w) ∈ E. Sometimes an edge has a third component known as a weight or a cost [8].

2.2. Weight of Sentence

Weight of sentence is a value of a sentence to determine how the sentence plays important meaning in a paragraph of an article. Clearly, higher weight of a sentence in a paragraph means existence of the sentence plays more important role in the paragraph. Thus, in the process of summarization, higher weighted sentences should have higher priority to be chosen as a part of summary. It is necessary to ignore unimportant words in the article before starting to calculate the weight of sentences. In other words, *stemmer process* and *stopword removal* have to be already implemented in the article before calculating the weight of sentences. Here, the weight of sentences consists of the following 4 component values, W1, W2, W3 and W4 based on [7] by some modifications as follows.

W1 be a score concerning similar words in a sentence compared with a list of keywords in the article. Supposed an article has a keywords list. Words in the sentence are compared to the words in the keywords list of the article. If there are more words in the sentence are similar to the keywords, the value of W1 will be higher.

W2 be a value to express the frequency of words of a sentence in a article. The number of same words in the sentence and the article is calculated. The result will be divided by the total words in the article by also considering the frequency of the words in the article. If the sentence has higher score result from the calculation, it means the sentence consists of more words that have high frequency in the article.

W3 be a value that is determined by the position of a sentence in a paragraph. Weight of sentence is also determined by the position of the sentence in a paragraph of an article. In general, every paragraph in a good writing of an article usually only provides one main idea. Related to the *Deductive-Inductive method* in the Indonesian grammar, the first and the last sentences in a paragraph are usually considered as main

sentences that express main idea of the paragraph. There are four location to put main sentence [1]:

a. At beginning of paragraph (Deductive Paragraph)
b. At end of paragraph (Inductive Paragraph)
c. At beginning and end of paragraph (Deductive-Inductive Paragraph)
d. No main sentence

Thus, we should consider giving a higher value to the first and the last sentences of every paragraph compared to the other sentences.

W4 be a weight to express relation between a sentence and the other sentences in the article. Here, this process of calculation is related with article mapping in which we need to calculate the number of relation (edges) of every sentence in the article. More number of relations from a sentence to the others in the article, the value of W4 will be higher that means the sentence is considered being more important in the article because the sentence probably discuss about article's main topic.

The result from each component values above will be added together and defined as a weight of sentence. Formally, let a paragraph has *n* sentences as given by $\mathbf{P}=\{S_1, S_2, \ldots, S_n\}$, where S_1 and S_n are the first and the last sentences, respectively. $W(S_j)$ is defined as a weight of a sentence S_j as simply given by the following equation:

$$W(S_j) = W1(S_j) + W2(S_j) + W3(S_j) + W4(S_j), \quad (1)$$

where $j \in N_n$.

2.3. Weight of Relation

A *weight of relation* between two sentences in a paragraph is similar to a cost or distance between those sentences. Consequently, if weight of relation between two sentences is less then the distance between two sentences is closer. Formally, let a paragraph has *n* sentences as given by $\mathbf{P}=\{S_1, S_2, \ldots, S_n\}$. Weight of relation between two sentences, S_i and S_j, is given by $R(S_i,S_j)$ as follows.

$$R(S_i,S_j) = \begin{cases} \dfrac{(j-i)^2}{\alpha(S_i,S_j) \times W(S_j)}, & i < j \\ \infty, & i \geq j \end{cases} \quad (2)$$

where $i, j \in N_n$. $\alpha(S_i,S_j)$ is defined by the number of similar words between S_i and S_j ignoring *stopword* in those sentences. Let S_i and S_j be also considered as a set of words in the sentences, S_i and S_j, respectively. $\alpha(S_i,S_j)$ is given by:

$$\alpha(S_i,S_j) = |S_i \cap S_j| \quad (3)$$

From (2), it can be verified that if $\alpha(S_i,S_j) = 0$ then $R(S_i,S_j) = \infty$ that means there is no relation between S_i and S_j. Suppose that $R(S_i,S_j)$ means a relation from S_i to S_j, then (2) consider that there is no relation from S_i to S_j if and only if $i \geq j$.

3. System Design

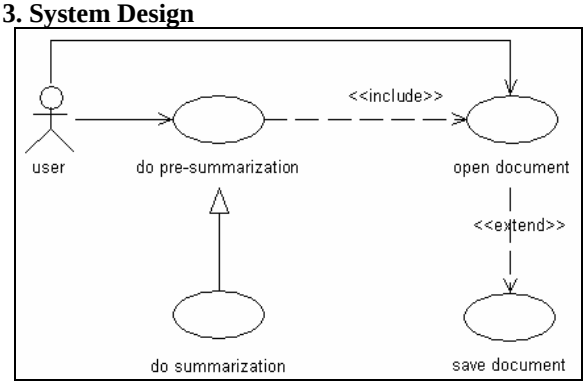

Figure 1. Use Case Diagram

There are 4 processes in the development of IATS (Indonesian Automated Text Summarization) as implemented in [10]:

1. **Open Document.** This process is to open the article and convert it to text document.
2. **Pre-Summarization.** In this process we separate the sentences in the document and process it to become collection of words. Next step in this process is *stemmer* process, to change back the words to basic form. The last step is *stopword removal* that is for eliminating unimportant words, e.g., conjunction, pronoun etc. For the implementation of this process, we use pre-processing document method that has been developed for another application [6]. This pre summarization process is useful to calculate weight of each sentence in article (see Section 2.2).
3. **Summarization.** This process is the main process for the implementation of IATS. In this process we summarize the article by firstly calculating weight of sentences as given in Section 2.2 and weight of relation as defined in Section 2.3. In this case, we used the concept of *weighted directed graph* to present relation between sentences in the paragraph of the article, in which every vertex is defined as a single sentence, and every edge expresses relation between two sentences in the paragraph. Here, the cost or weight of every edge is given by *weight of relation* as described in Section 2.3. The process of summarization is simply given by a shortest path from the first sentence to the last sentence using *Dijkstra's Algorithm* as well as *Steepest Ascent Hill Climbing Algorithm*. For example, given a paragraph of an article have five sentences. Relation of all sentences is represented in a weighted directed graph as shown in Figure 2. The

graph consists of five vertexes representing five sentences.

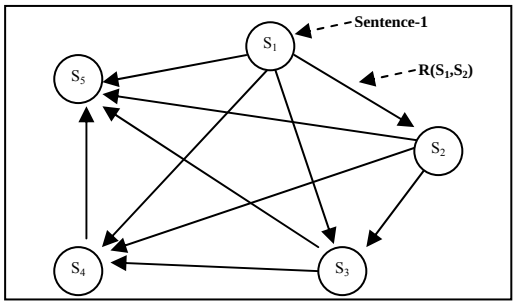

Figure 2. Weighted-Directed Graph

Next step in this process is to change back the article to the original sentences, but eliminating the sentences which are not listed in the path list. The result of summarization will be shown to the user.

4. **Save Document**. In this process, we save the summarization result in Ms-Word format (*.doc).

The activity diagram for all processes above can be seen in Figure 3 until Figure 6.

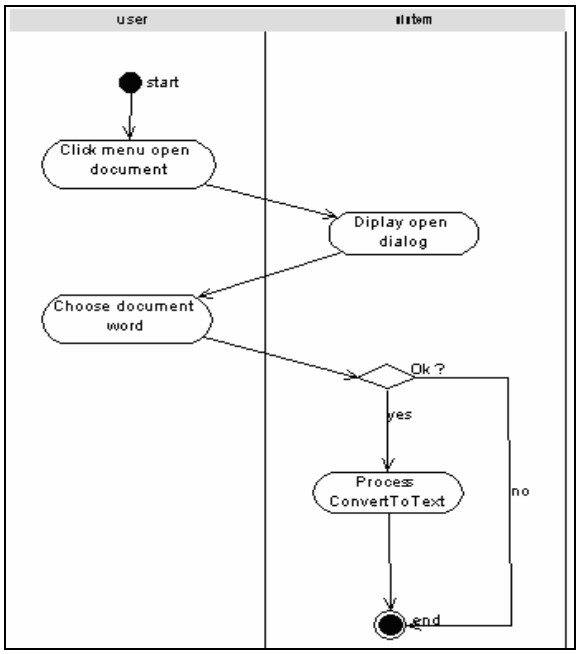

Figure 3. Open Document Activity Diagram

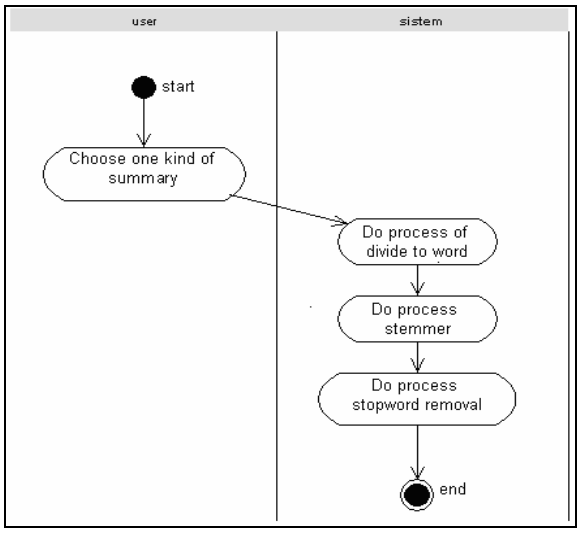

Figure 4. Pre-Summarization Activity Diagram

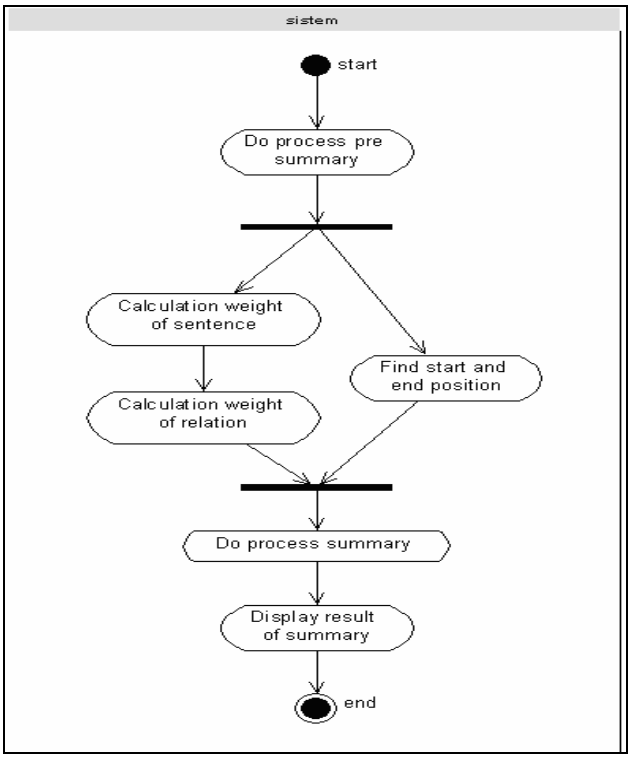

Figure 5. Summarization Process Activity Diagram

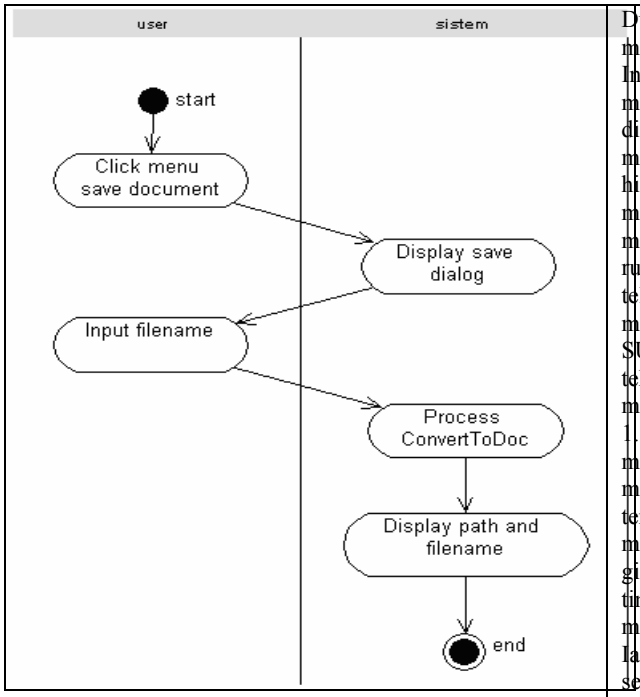

Figure 6. Save Document Activity Diagram

4. Experimental Result

In this section we show the experiment results that had been already implemented in [10]. Results of examinations are given as follows:

The specification of hardware and software we used for experiment is:

- Processor Pentium IV 1600 MHz
- Memory 512 Mbyte
- HardDisk 40 Gigabyte
- Operating System Windows XP Professional
- Database Ms. Access 2003

Figure 7 shows an original article and its result of summarization is given in Figure 8. And the result of some experiments is shown in Table 1.

Pengaruh SUTET Terhadap Kesehatan

Dr Anies, peneliti Saluran Udara Tegangan Ekstra Tinggi (SUTET) dari Fakultas Kedokteran Universitas Diponegoro Semarang menegaskan, sampai saat ini pengaruh medan listrik SUTET terhadap kesehatan manusia masih kontroversial, meski dari berbagai riset yang dilakukan, muncul keluhan, seperti mual, pusing, hingga sulit tidur. "SUTET belum sampai menimbulkan gangguan kesehatan permanen. Kalau keluhan seperti itu memang ada, tapi sekali lagi, itu bukan penyakit," kata Anies ketika diminta tanggapan di Semarang, Ahad malam sehubungan makin maraknya tuntutan ganti rugi dari korban SUTET di Jawa Tengah. Anies yang pada tahun 2004 melakukan penelitian dampak SUTET di Tegal, Pemalang, dan Batang tersebut mengakui, memang muncul keluhan pada warga yang tinggal di sekitar SUTET, namun tidak sampai menimbulkan gangguan kesehatan serius.

Ia menyebutkan, menurut standar Badan Kesehatan Dunia (WHO), medan listrik di bawah SUTET maksimum lima kV/meter, sedangkan Ikatan Dokter Indonesia pada tahun 1997 menetapkan ukuran medan magnit maksimum 0,1 miliTesla (mT). Anies yang disertasi doktornya berisi penelitian SUTET itu menegaskan, SUTET yang ada di sepanjang Jawa Barat hingga Jawa Timur masih berada di bawah batas maksimum standar WHO maupun IDI. Menurut dia, medan listrik jauh lebih besar justru berada di dalam rumah, misalnya pesawat televisi, monitor komputer, telepon seluler hingga "microwave" yang memiliki medan listrik berjuta kali lipat dibanding medan listrik SUTET. Ia menyebutkan, telepon seluler (HP) pada awal teknologi seluler ini ditemukan memiliki kekuatan 900 megaHertz, tetapi sekarang dua kali lipat yaitu rata-rata 1.800 megaHertz dan 1.900 megaHertz. "Bandingkan medan listrik di bawah SUTET yang masih di bawah 50 megaHertz," katanya. Medan listrik jauh lebih tinggi lagi terdapat pada "microwave" yang radiasi panasnya menimbulkan medan listrik hingga 2,45 giga Hertz. Satu giga Hertz sama dengan satu miliar Hertz. Dengan tingginya medan magnet tersebut masyarakat tidak mempermasalahannya karena mereka membutuhkannya. Ia menyarankan, untuk mengurangi efek SUTET, di sekitar areal SUTET ditanami pohon-pohonan sehingga radiasi listrik berkurang.

Figure 7. Original Article "Pengaruh SUTET Terhadap Kesehatan"

Pengaruh SUTET Terhadap Kesehatan

Dr Anies, peneliti Saluran Udara Tegangan Ekstra Tinggi (SUTET) dari Fakultas Kedokteran Universitas Diponegoro Semarang menegaskan, sampai saat ini pengaruh medan listrik SUTET terhadap kesehatan manusia masih kontroversial, meski dari berbagai riset yang dilakukan, muncul keluhan, seperti mual,pusing ,hingga sulit tidur." SUTET belum sampai menimbulkan gangguan kesehatan permanen. Kalau keluhan seperti itu memang ada, tapi sekali lagi, itu bukan penyakit," kata Anies ketika diminta tanggapan di Semarang, Ahad malam sehubungan makin maraknya tuntutan ganti rugi dari korban SUTET di Jawa Tengah. Anies yang pada tahun 2004 melakukan penelitian dampak SUTET di Tegal, Pemalang, dan Batang tersebut mengakui, memang muncul keluhan pada warga yang tinggal di sekitar SUTET, namun tidak sampai menimbulkan gangguan kesehatan serius.

Ia menyebutkan, menurut standar Badan Kesehatan Dunia (WHO), medan listrik di bawah SUTET maksimum lima kV / meter, sedangkan Ikatan Dokter Indonesia pada tahun 1997 menetapkan ukuran medan magnit maksimum 0,1 miliTesla(mT). Dengan tingginya medan magnet tersebut masyarakat tidak mempermasalahannya karena mereka membutuhkannya. Ia menyarankan, untuk mengurangi efek SUTET, di sekitar areal SUTET ditanami pohon-pohonan sehingga radiasi listrik berkurang.

Figure 8. Summarization Result of Article "Pengaruh SUTET Terhadap Kesehatan"

Table 1. Experimental Result of Summarization and Processing Time

Num	Articles			Summarization Results							
	Titles	Paragraphs Counts	Words Counts	Words Count		Words Differences (%)		Relation Values (%)		Processing Time	
				Dijkstra	SAHC	Dijkstra	SAHC	Dijkstra	SAHC	Dijkstra	SAHC
1	Bagaimana Terjadinya Kangker	26	1294	998	964	23	26	67	65	01m00s718ms	00m03s328ms
2	Belajar Music Perkuat Daya Ingat Si Kecil	6	271	270	270	0	0	85	86	00m04s516ms	00m00s843ms
3	Khasiat Teh Hijau	9	359	302	302	16	16	73	73	00m06s953ms	00m02s141ms
4	Lagi, AS pasok Bom Ke Israel	16	982	482	471	51	52	41	40	03m00s735ms	00m04s218ms
5	Media Saat Ini Merupakan Refreksi Dari Krisis Di Fiji	31	1934	1605	1523	17	21	75	72	00m30s094ms	00m06s000ms
6	**Pengaruh SUTET Terhadap Kesehatan**	**3**	**305**	**177**	**149**	**42**	**51**	**58**	**47**	**00m45s109ms**	**00m00s734ms**
7	Aladin Dan Lampu Ajaib	9	813	317	296	61	64	31	29	06m59s922ms	00m02s156ms
8	Petualangan Sinbad	8	781	254	240	67	69	25	24	04m51s250ms	00m02s672ms
9	Si Kancil Kena Batunya	8	539	254	250	53	54	38	38	03m36s407ms	00m01s375ms
10	Kebakaran Hutan	23	700	515	695	26	1	85	85	00m06s141ms	00m03s531ms
11	Surat MenKeu Ditarik	24	764	760	760	1	1	92	92	00m03s250ms	00m02s406ms
12	Sangkuriang	6	397	169	169	57	57	37	37	01m28s140ms	00m01s093ms
13	Jack Dan Pohon Kacang	13	984	348	343	65	65	28	28	06m39s078ms	00m02s812ms
14	Wanita Lebih Mudah Bete Di Pagi Hari	6	329	224	224	32	32	59	59	00m02s750ms	00m01s484ms
15	Penemuan Planet Baru	11	595	492	493	17	17	73	73	00m11s047ms	00m02s594ms

Notes:
Djikstra : Summarization by using Exhaustive Shortest Path Dijkstra's Algorithm.
SAHC : Summarization by using Heuristic Shortest Path Steepest Ascent Hill Climbing Algorithm.

In the Table 1, we also compared the result from experiment using *Heuristic Stepeest Ascent Hill Climbing Algorithm* and *Dijkstra's Algorithm*. The main goal using Heuristic Stepeest Ascent Hill Climbing Algorithm instead of Exhaustive Dijkstra's Algortihm is to save summarization processing time.

4.1 Experimental Result by Interview

This experiment is to test how much the summarization result can make the reader understand about the whole original article contents. Interview is done to 2 people by giving the summarization result and some questions. Those questions are created using original article without look at the summarization results and the topics in the question are about main topics of the article.

For the first article "Bagaimana Terjadinya Kanker", the list of questions are:

1. Apa yang dimaksud dengan kanker ?
2. Apa yang membentuk sel-sel kanker ?
3. Apa saja faktor resiko terjadinya kanker ?
4. Apa hubungan kanker dengan riwayat hidup keluarga ?
5. Virus apakah yang menyebabkan terjadinya kanker leher rahim pada wanita ?

And for the second article "Kebakaran Hutan", the list of questions are:

1. Dimanakah terjadinya kebakaran hutan di daerah Sumatera ?
2. Kapan terjadinya kebakaran tersebut ?
3. Apa akibat dari kebakaran di Aceh ?
4. Bagaimana cara memadamkan api tersebut ?
5. Berapa perusahaan yang menjadi tersangka pembakaran hutan ?

From five questions, the two respondents cannot answer only one question. This is because the summarization result didn't contain the answer of the question.

5. Conclusions

In general, the proposed application gave a better result for argumentative articles compare than descriptive articles, because in descriptive article, the summarization result can loose important sentences and some relations between sentences. In the argumentative articles, the system can summarize well and in general can cover almost all main topics in those articles.

Exhaustive Dijkstra's Algorithm is good for the readers who consider the importance of summarization result compare to the processing time.

In general, a more number of words and sentences in a paragraph will provide a shorter result of summarization compare to the original article.

References

[1] Akhadiah, Sabarti, M.K.; Arsjad, Maidar; Ridwan, Sakura. (1986). *Buku Materi Pokok : Bahasa Indonesia*. Jakarta: Penerbit Karunika Jakarta UT.

[2] Cormen, H. Thomas; Leiserson, E. Charles; Rivest, L. Ronald. (1990). *Algorithm*. London : The MIT Press.

[3] Intan, Rolly; Defeng, Andrew. (2005). *HARD : Subject-Based Search Engine Menggunakan TF-IDF Dan Jaccard's Coefficient*. Surabaya : UK Petra.

[4] Keraf, Gorys. (1984). *Tata Bahasa Indonesia*. Jakarta : Nusa Indah.

[5] Russell, J. Stuart ; Norvig, Peter. (1995). *Artificial Intelligence – A Modern Approach*. New Jersey : Prentice Hall.

[6] S. Budhi, Gregorius; Gunawan, Ibnu; Yuwono, Ferry. (2007). *Algoritma Porter Stemmer For Bahasa Indonesia Untuk Pre-Processing Text Mining Berbasis Metode Market Basket Analysis.* Jurnal Pakar.

[7] Sjobergh, Jonas; Araki Kenji. (2005). *Extraction Based Summarization Using A Shortest Path Algorithm*. Sweden: KTH Nada.

[8] Weiss, Mark Allen. (1994). *Data Structures And Algorithm Analysis in C++*.The Benjamin/Cummings Publishing Company, Inc.

[9] Zaenak, Arifin ; Tasai, Amran. (2004). Cermat *Berbahasa Indonesia Untuk Perguruan Tinggi*. Jakarta: Penerbit Akademika Pressindo.

[10] Stevanus R.,R. (2006), *Perancangan dan Implementasi Automated Text Summarization Menggunakan Algoritma Exhaustive dan TF-IDF*, Undergraduate Thesis, UK.Petra.