Generating Indicative and Informative Summaries
For Search Engine Results

Yudi Wibisono¹, Dwi Hendratmo Widyantoro*²

¹School of Electrical Engineering and Informatics –Bandung Institute of Technology,
Jl. Ganesha 10 Bandung

In internet search engine, webpage summaries can help users to quickly decide which webpage to read from a list of search results, reducing the time and effort to visit irrelevant webpages. We create a search result presentation system that combines both indicative summary and informative summary. Indicative summary facilitates a quick scanning among the search results, and informative summary provides succinct description of the original document. This paper discusses the details techniques we develop to generate indicative and informative summaries.

1. Introduction

Typical search engine returns ranked list of webpage to respond user query. Often it gives very long list of recommended webpage and users have to choose the most relevant document. In ideal system, user shouldn’t have to refer full version of webpage to make this decision. Webpage summarization can help user in this task by provide shorter version of original document.

Webpage summarization has recently gained a lot of attention from researchers in the field as this topic offers a promising method to address a variety of problems arising from the proliferation of information on the Internet.

Summarization is a process that compress original document by extracting the most important parts of document. Based on its function, there are two kinds of summarization: indicative and informative summarization.

Indicative summarization results two to three lines summaries that suggest the contents of original documents closely related to a user query. This summary facilitates a quick scanning among the search results.

Informative summary is meant to represent the original document, providing an idea of what the whole content of document is all about. Current major search engines (Google, Yahoo, MSN, ask.com), however, only utilize indicative summary as the basis for users to decide whether or not to visit the corresponding webpage. Although quite effective, indicative summary is not accurate enough to predict the document content, occasionally misleading users to visit wrong websites. The informative summary addresses this problem by providing more accurate information about the document content.

This paper will present use of indicative and informative summarization for search engine results.

The rest of this paper is organized as follows: Section 2 discusses system overview. Summarization technique used in our system is explained in Section 3. Summarization result is discussed in Section 4. The last section gives concluding remarks and future work.

* Email: dwi@if.itb.ac.id

2. System Overview

Figure 1 shows screen snapshot of our system. After a user gives query to search engine, results will be presented as ranked list of web pages that consist of title and indicative summary.

If the user needs more information about the document, he/she can get informative summary of the document simply by hovering mouse pointer to word “More” located next to indicative summary.

Informative summary is generated through sentences selection from a document. Each sentence is scored based on the values of its features such as words and sentences location. The summary for each document is then generated from the top-scoring sentences, until a desired summary length is reached. Indicative summary is generated similarly, but is biased by user query. Therefore, informative summary of a document is constant, but indicative summary depends on user query.

Sentence scoring and generation of informative summary is conducted offline when the search engine builds its index (see Figure 2). Sentence score is stored because we will use it later for building indicative summary. We also store informative summary because its value is constant.

On the other hand, indicative summary is generated on-the-fly. It is generated right after a user entered a query for searching webpage (see Figure 3) by using query score and other sentence scores. Search engine will use index to search webpage and then display search results with indicative summary (and informative summary when user needs it).

The length of summary generated by the system is adjustable. There are two parameters to adjust summary length: compression ratio and maximum number of words in summary. For example, if the compression rate is set to 0.1, the summary length will be 10% of original document. As default value for informative summary, we set compression ratio to 0.2, and for indicative summary, the summary length is limited to 20 words. We use Lucene [5] as search engine tools.
We use sentence extraction technique to generate both informative and indicative summaries. Sentence extraction is chosen because we need robust and efficient summarization. Sentence extraction result is also more readable because it uses sentence from original document.

The algorithm for generating summarization proceeds as follows:
1. Pre-process input document (case folding and stop word removal)
2. Parse document into sentences.
3. Calculate sentence score.
4. Sort sentences based on score
5. Generate summary by adding top-scoring sentences until desired summary length reached.
6. Resort sentences order in summary based on order in original document.

We use four aspects to calculate sentence score [1,2]:

1. Cluster of significant word.
2. Title Term Frequency
3. Location/Header
4. Query biased (only for indicative summary).

### 3.1. Significant-Word Cluster Score

This technique is based on Luhn’s word clustering [4] and, which is then modified by Lam-Adesina [1]. A word is called significant in a document if the document contains that word at least with certain frequency.
Minimum word frequency is defined as follows:

\[
\begin{align*}
\text{If } (ns < 25) & \quad \text{min}_\text{freq} = 7 - (0.1 \times (25 - ns)) \\
\text{Else if } (ns > 40) & \quad \text{min}_\text{freq} = 7 + (0.1 \times (ns - 40)) \\
\text{Else} & \quad \text{min}_\text{freq} = 7
\end{align*}
\]

where \( ns \) is the number of sentence in document.

Commons word like “the”, “and”, “are” are removed in pre-processing stage. After all significant words in a document are found, system searches for clusters of significant word. Cluster is formed by significant words if they are separated by less than five non significant words.

For example, \( w_1, w_{i+1} \ldots w_n \) is word sequence. There are 14 words (word in bold is a significant word):  \[ w_1 \ w_2 \ w_3 \ w_4 \ w_5 \ w_6 \ w_7 \ w_8 \ w_{10} \ w_{11} \ w_{12} \ w_{13} \ w_{14}. \] Because \( w_2 - w_4 - w_7 \) is significant word and separated by less than five non significant word, then \( w_2 - w_4 - w_7 \) become one cluster (represent by bracket). But although \( w_{14} \) is significant word, it is separated by 6 words from \( w_7 \) so cluster can not be expanded to \( w_{14} \).

Cluster score is given by Eq (1):

\[
CS = \frac{SW^2}{TW}
\]

where \( CS \) = Cluster Score, \( SW \) = Number of significant word in cluster, \( TW \) = Total number of word in cluster.

If there is more than one cluster in a sentence, we use cluster with the biggest cluster score.

3.2. Title Score

Title score is calculated based on how many title word appear in sentences. Title score is computed as follows:

\[
T_S = \frac{TTS}{TTT}
\]

where \( T_S \) = title score \( TTS \) = number of title word found in sentence. \( TTT \) = number of word in sentence.

3.3. Header Score

Edmunson (3) found that position of sentence in the document can be used to determine the importance of sentence. Sentence located on first paragraph is assumed more important than other sentence on other location.

Sentence score is applied only for two first sentences and computed as follows.

3.4. Query Biased Score

Query biased score is used only in indicative summary generation. Like title score, it is based on how many query terms are found in a sentence. With query score, indicative summary for a document can vary based on query. Query score is given by Eq (4):

\[
QS = \frac{TQ^2}{NQ}
\]

where \( TQ \) = number of query term in sentence. \( NQ \) = number of query term.

3.5. Combining Score

Sentence score in informative summary generation is computed by adding the first three aspect as follows:

\[
\text{Sentence Score} = \text{Cluster} + \text{Title} + \text{Header}
\]

Indicative summary generation use all aspects.

\[
\text{Sentence Score} = \text{Cluster} + \text{Title} + \text{Header} + \text{Query}
\]

4. Summarization Result

We had applied our system to several news webpage. Although we haven’t evaluated it formally the initial result is promising. The following example shows indicative summary, informative summary and original document.

**Indicative Summary (query: “China”):**

China is Japan's largest trading partner, and Japan is China's third-largest trade partner after the European Union and the U.S.

**Informative Summary:**

Trade between China and Japan has almost quadrupled in the last decade, reaching $211 billion last year. China is Japan's largest trading partner, and Japan is China's third-largest trade partner after the European Union and the U.S. During Wen's visit, which begins Wednesday and ends Friday, China is expected to lift a 4-year-old ban on Japanese rice imports, and Japan will offer to transfer energy-saving technology to China.

**Original Dokumen:**

**Title:** China, Japan to meet in first visit since 2000

**Original Content:**

BEIJING -- Chinese Premier Wen Jiabao will fly to Japan this week to mend relations between neighbors tightly linked by trade but torn by years of warfare and squabbling.

On the first such visit by a Chinese premier since 2000, Wen will make a rare speech to Japan's parliament, confer with Prime Minister Shinzo Abe and meet Emperor Akihito. He also will seek to assure
widespread concerns among Japanese that China's rise threatens the region. Tensions have subsided markedly between the neighbors since Abe came to office in September, replacing Junichiro Koizumi, whose regular visits to Yasukuni Shrine -- which honors Japan's war dead, including war criminals -- enraged China. Abe has not visited the shrine while in office.

In a sign of the unfolding rapprochement, China responded with restraint to Abe's denials in early March that Japan's military had forced Asian "comfort women" into military brothels during World War II, a lightning-rod issue elsewhere in East Asia.

Leaders of both nations appear determined not to let an array of current and historical disputes get in the way of thriving economic relations.

Trade between China and Japan has almost quadrupled in the last decade, reaching $211 billion last year. China is Japan's largest trading partner, and Japan is China's third-largest trade partner after the European Union and the U.S.

During Wen's visit, which begins Wednesday and ends Friday, China is expected to lift a 4-year-old ban on Japanese rice imports, and Japan will offer to transfer energy-saving technology to China. The two nations will launch a ministerial-level economic dialogue, raising ties to a strategic level.

5. Concluding Remarks and Future Work

In this paper, we have shown that indicative and informative summaries can be used to present search results. Indicative summary characteristic is different from informative summary. Indicative summary is much shorter than informative one. Hence it may need different method. Currently, we use news webpage for experiments. In the future, we plan to apply it to other kind of webpage. We also plan to do formal evaluation using task based method.

References


