# UpdateNews: A News Clustering and Summarization System Using Efficient Text Processing

Takaharu Takeda
The Graduate University for Advanced Studies
2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan
takeda_takaharu@grad.nii.ac.jp

Atsuhiro Takasu
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan
takasu@nii.ac.jp

## ABSTRACT

This paper proposes a news article clustering and summarization system. It provides integrated and effective access to news articles from various news cites. Proposed system consists of a crawler, topic detector and summarizer of news articles. This paper focuses on its efficient summarization technique to handle large amount of crawled news articles.

## 1. INTRODUCTION

Nowadays various kinds of information sources are accessible through the Internet. News cites are one of the useful information sources among them. Because they are distributed and frequently updated, we need a new technology that integrates dynamically changed documents.

In order to provide news articles effectively, a system should provide various kinds of functions. First, users want to read only interested news articles. Therefore, the system should help users to select news according to their preferences. Categorization of news articles is first step to this problem. Many news cites provide articles by listing them according to categories such as politics and sports. However, the granularity of categories are too coarse to filter articles. The topic detection and tracking technology [1] detects events from news articles and makes clusters of articles according to the event. With this technology users can select and keep track of interested news event.

Because same event can be described from different aspects in news articles, users often compare articles from different sources. Therefore, the system needs to gather news articles from various sources and link articles describing the same topic. Google News [1] is an example of this function.

Usually the news articles are updated frequently and their description are overlapped each other in series of news articles. By digesting the news articles and removing the duplicated description, users can obtain the information efficiently. Newsblaster [2] provides summarized text of the news articles on the same topic. This paper proposed our news article clustering and summarization system [3]. It is similar to Newsblaster system [2], however it provides efficient summarization function. As a result. it can quickly process the clustering and summarization of news articles.

[1]http://news.google.com/
[2]http://www.newsblaster.com/


*JCDL* '07 Vancouver, Canada


## 2. SYSTEM OVERVIEW

Figure 1 shows the entry page of the system that gives an overview of current news topics. Currently our system uses 8 categories which are given from source news cites. Extracted topics in each news category are listed with their head lines. When selecting one topic, the system gives the summary of the topic with the list of news articles.


**Figure 1: System Overview.**

This system consists of three modules: a crawler, a topic detector and a topic summarizer. The crawler gathers news articles. It accesses a web cite [4] that keeps track of news update of major news cites and obtains the information about uploaded news in RSS formats. Then it accesses the original news cite and obtains new articles.

For the crawled news articles, our system assigns a topic. There are many studies on the topic detection and tracking [1]. In this system we use a simple topic detection method based on the vector space model. Each article is represented with a term vector weighted by the inverse document frequencies. Then, we measure the similarity between the documents by cosine measure that is frequently used in information retrieval. Let a topic consist of a set $A$ of news article. Then the similarity between the topic and a news article $a$ is measured as

$$\mathrm{sim}(A, a) = \max_{b \in A} \mathrm{dsim}(a, b)$$

[3]http://updatenews.sub.jp/
[4]http://news.ceek.jp/

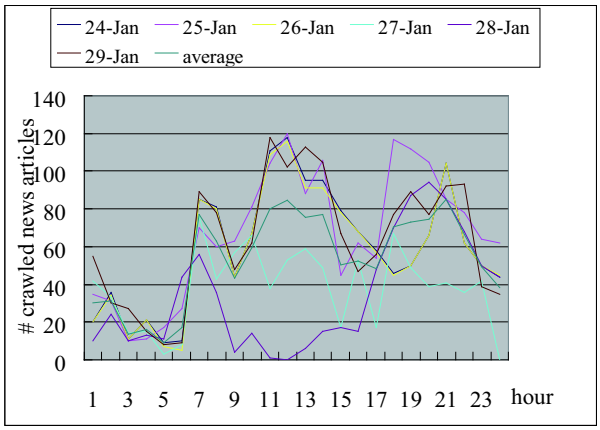


**Figure 2: Number of Crawled News Articles.**

where $\mathrm{dsim}(a, b)$ denotes the cosine measure similarity of articles $a$ and $b$. We search the most similar topic and if the similarity is more than a threshold, the news article $a$ is assigned to the topic. Otherwise the news article $a$ is assigned to a new topic.

We describe the summarization method in the next section. Since June 2006, we have been running this system and it contains 8240 articles that are clustered into 354 topics as of February 2, 2007.

## 3. NEWS CLUSTER SUMMARIZATION

Figure 2 shows the number of news articles crawled per hour which is obtained from the system log during Jan. 23 to 31. As the graph shows the number of crawled articles has a peak around noon and the maximum number of articles is about 120 per hour. In order to process this volume of articles, we need efficient clustering and summarizing methods. In our system, since summarization is the dominant part of the news article processing, we developed an efficient summarization method using a light text processing.

For summarization, we need to remove sentences that are similar to others in articles in a topic. Our system uses edit distance for measuring the similarity between sentences where edit costs are determined based on word weight given by inverse document frequencies. Precisely. for each word $w$ in sentences, cost of both insert and delete $w$ is defined as

$$C_i(\lambda, w) = C_d(w, \lambda) = \log |D| - \log df(w)$$

where $df(w)$ denotes the number of documents containing the word $w$ in the topic and $|D|$ denotes the number of documents in the topic. This means that we can remove or insert a word with low cost if it appears frequently in the topic like stop words. Using the costs of insert and delete operations, the substitution cost is defined as

$$C_s(w_1, w_2) = C_i(\lambda, w_1) + C_d(w_2, \lambda) \quad .$$

For a word sequence $\mathbf{w}_1$ and $\mathbf{w}_2$ whose lengths are more than a threshold $l$, if the edit distance of them is less than a threshold $\sigma$, we call $\mathbf{w}_1$ and $\mathbf{w}_2$ are similar word sequences.

For a set $D$ of news articles belonging to a topic, we make summary by the following procedure:

1. select the latest 10 news articles in $D$,

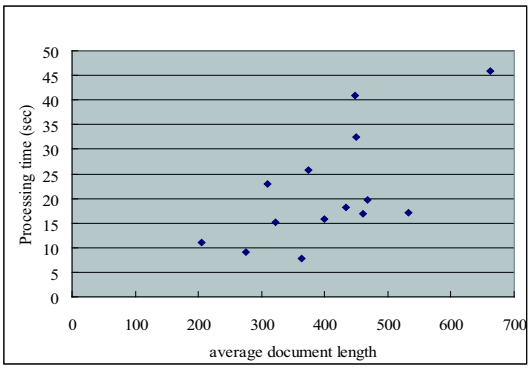


**Figure 3: Processing Time for Summarization.**

2. choose the *seed article* $d$ satisfying the condition:

$$\arg\max_{d \in D} \sum_{e \in D} \mathrm{dsim}(d, e) \quad ,$$

   and choose a *similar article* $e \in D$ that is the most similar to the seed article,

3. enumerate pairs of similar word sequences $\mathbf{w}_1 \in d$ and $\mathbf{w}_2 \in e$, and compose a *seed summary* consisting of sentences containing one of the similar word sequences,

4. For each sentence in the latest 10 news articles, if it does not contain any word sequence that is similar to word sequences in the seed summary, add the sentence to the seed summary.

In this procedure, the enumeration of similar word sequences requires the highest computational cost. We can solve this problem using the suffix array data structure. Figure 3 shows the processing time w.r.t. the average document length. As graph shows we need more time for longer articles and average processing time is about 20 seconds that is efficient enough to process crawled articles on demand.

## 4. CONCLUSIONS

This paper proposes a news articles clustering and summarization system. The proposed system realizes efficient news topic summarization. Generated summaries looks good. We plan to assess the quality of summarization by comparing with other methods. Because the proposed summarization method is language independent in nature. we plan to apply the method to multi-language news articles.

## 5. REFERENCES

[1] J. Allan. *Topic Detection and Tracking: Event-based Infromation Organization*. Kluwer, 2002.

[2] K. R. McKeown, et al. Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In *Proceedings of the Human Language Technology Conference*, 2002.