Hierarchical Summarizing and Evaluating for Web Pages

Kou TAKAHASHI[1], Takao MIURA[1], and Isamu SHIOYA[2]

[1] HOSEI University, Kajinocho 3-7-2, Koganei, Tokyo, JAPAN
[2] Sannno University,Kamikasuya 1563, Isehara, Kanagawa, JAPAN

Abstract. In this investigation we propose a novel summarization method of Web pages using hierarchical expression. We discuss close relationship between summarization and hierarchical clustering to obtain the results, and we examine how to evaluate hierarchical summarization based on both correlation and structural aspects. We describe some experimental results using NTCIR Web documents to examine our method.

1 Introduction

Nowadays we face to huge amount of Web pages which keep growing everyday. Whenever we like to know current situation of interests, we can examine easily what's going now all over the world through these pages. However, at the same time, such information flooding makes it impossible for us to grasp the contents quickly and intuitively. That's reason why much attention has been paid on how to grasp and summarize the contents quickly [7, 4]. To attack these issues, there have been several approach proposed, and among others, we have two important techniques, *clustering* and *summarization*[12, 4].

Clustering means a method to put objects into several groups in such a way that objects in a group are similar with each other while objects in distinct groups are not [6]. In other words, clustering depends on both definition of similarity and algorithm by which we can extract interesting patterns that are hidden. Most information in Web pages are categorical (say, we see "Computer Science", "Biology", "Mathematics") but not numerical so that we interpret hardly the notions of metric or order. This is the true reason traditional clustering techniques are not well-suited.

Summarization provides us with presentation of important contents of information source in a simplified way by which we can grasp quickly. In investigation of automatic summarization, two major aspects have been discussed, *extraction* and *abstraction*. To extract important part from the contents, we put weights such as frequency to words or paragraphs, and score them. Eventually we select sentences with high score order. On the other hand, to abstract contents, we should analyze them in terms of Natural Language Processing (NLP) or Information Retrieval (IR), thus few approaches has been proposed so far.

In this work, we propose a new kind of Web summarization technique based on hierarchical clustering. there are two kinds of ideas, Web page clustering and Web page summarization. We put our attention on *frequency* and *co-occurrence* of both index words and hyper-links appeared in the Web pages. We decompose the page contents into semantic textual units (STU) as described later.

With the help of hierarchical clustering among them, we extract some sort of structure among collections of STUs so that we can consider the structure as the abstraction of the contents in a form of hierarchy. We obtain a center STU from each cluster, thus we can represent the summarization in terms of structure among STUs.

We must discuss how to evaluate the hierarchical summarization. Generally evaluation is difficult as to both clustering and summarization, This is reason because *correct* output can't be defined. In this investigation we propose a novel evaluation method, from view points readability of both clusters and hierarchical structures as well as reading comprehension. We evaluate the hierarchical summarization in terms of penalty, called *cost*.

In section 2 we introduce several issues of automatic summarization and application to Web, and we discuss a new technique how to summarization the contents in a hierarchical manner in section 3. In section 4 we propose a novel evaluation method of hierarchical expression. We show some experimental results in section 5, and conclude our work in section 6.

2 Automatic Summarization

In this section we review automatic summarization putting our attention on Web pages[7]. Generally *summarization* means a process to distill most important information from document sources for particular users and tasks. The automatic summarization for particular users is used for contents grasping of documents(such as news articles). There are 3 kinds of the techniques proposed, *Extracting*, *Abstracting* and *Clustering*.

Extracting means identifying the most relevant parts to the main theme appeared within the document. This approach is advantageous because very often statistical techniques such as frequency and correlation correspond to the relevance or importance. Here we don't need any background knowledge of NLP, and we could automate the process easily and efficiently. On the other hand, the results can be lack of coherence because of dangling anaphor such as pronoun and conjunction [7].

Abstracting means putting text objects into more general and super-ordinate concepts so that the results don't contain the expressions not appeared in the objects. The approach is advantageous because there should be coherent in the results, and better compression rate could be expected. On the other hand, we should analyze what documents mean to generalize context to some extent and NLP or IR technique should be required [7].

Clustering means collecting objects into several groups by using a certain similarity. We could extract underlying semantics from the objects. But we need other techniques such as labeling groups to interpret the results. The approach is advantageous because we can apply clustering easily and efficiently to objects without any NLP knowledge, but is difficult to define similarity and interpret the results, although several labeling have been proposed [6, 8].

As for some applications of automatic summarization for Web pages, extracting has been discussed for the purpose of hand-held device [3]. The technique is indispensable since the display is really small for Web browsing. Web pages are divided into small units, called *semantic textual unit (STU)*, where each STU is defined as a unit of meaningful contents, usually corresponded to paragraphs or caption parts. Every first line is sent for browsing purpose. For deeply reading, users may ask of first 3 lines or whole lines.

A different approach comes from Web clustering, since each cluster corresponds to closely related collection of Web pages with each other. Labeling clusters corresponds to summarizing the page contents, because the labels represent what the contents of the cluster mean. Very often frequent words could be seen as the labels of the pages [7]. But because there are strong similarities among the pages obtained by search engines, the frequency can't distinguish one from others [8]. Better approach could come from *important* words rather than frequent words. With these words we could see what's going on quickly. Moreover, there has been pointed out that labels consisting of sequences rather than the ones of words be helpful to grasp the contents. A typical approach as Suffix Tree Clustering (STC)[15] where sequences of words are used. However, STC is based on heavy assumption that we can use a set of documents which are closely related to each other(for example, hit list by search engine). Some investigation has proposed to extract important words based on Key-Graph [10] and word-sequences based on Suffix Tree [15], and then combined both results to abstract the pages [9]. However, the results depend on temporal aspects as well as clustering.

3 Summarizing Pages Hierarchically

Web documents consist of *tag* parts and *link* specifications as well as textual parts. We examine relationship over Web pages, and propose *structuring* as a method of novel automatic summarization. By structuring objects, we mean a technique to restate documents in terms of data structure. Since any data structure carries its own meaning, we use specific structure concisely and clearly as a substitute for parts of documents, thus we understand what the information does imply. Note that any data structure can be applied to any situation because they are polymorphic without any NLP knowledge.

However what kinds of structures are suitable for our situation and how to organize our documents appropriately ? What we work with are either words (minimum lexical units) or sentences (minimum semantic units). In the former case, we can examine fine semantics but hardly grasp global view because of detail relationship among words. On the other hand, in the latter case, it seems easier to examine relationship among sentences although outlines depend on syntax. Frequency and their correlation help us to obtain important parts for words, while distance between sentences and centroids of document clusters could show us what parts should be extracted. Thus, to grasp the contents of documents, important words and the relationship among them or document clusters play important role. There can be several expressions to describe semantic structure of the contents. Possible examples are directed graph and trees, since we expect abstraction mechanism to represent a variety of levels of the contents. In a case of tree, we think nodes in higher levels correspond to overview/global viewpoint while nodes in lower levels to detail/local aspects. One of these techniques is *Key*

Graph [10] by which the relationship can be modeled by means of graphs. However, there is no mechanism to interpret the contents from several abstraction levels of viewpoints.

In this investigation, we propose a novel technique for Web page summarization. One of the main issues is how to make groups to a huge amount of Web pages. It is well-known that naive clustering of Web pages generates a few gigantic clusters and many trivial clusters so that clusters provide us with nothing new knowledge. However, we have already discussed a sophisticated clustering method to Web documents based on correlation of hyperlinks with naive clustering under vector space modeling [12]. With these results, let us go one step further. In this work, we apply hierarchical clustering to each group of Web pages using STUs, and we discuss a new "labeling" method to each group by taking a centroid. Relationship among the clusters corresponds to the hierarchical structure as shown in the figure 1.

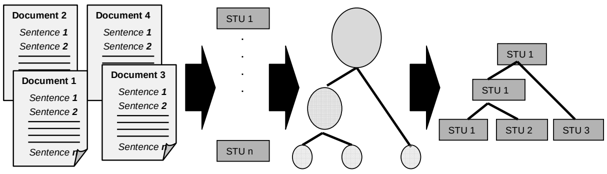


Fig. 1. Overview

3.1 Combination Clustering

In this section and the next section, we discuss how to make hierarchical summarization of Web pages. The process consists of two steps. The first step is that we put a collection of Web pages into disjoint groups by combining two kinds of clustering results, called *combination clustering*. The second step is that we summarize each group. The final results are a set of hierarchical expressions which can be seen as summarization of the Web page groups.

Our technique of combination clustering is made based on an idea that similar documents share many hyperlinks. We extract characteristic hyperlinks as well as index words to distinguish from each other. It is shown that the approach is effective to obtain fruitful results[12]. Here let us explain the outline of this approach by examples. Basically we put attention on co-occurrence of hyperlinks. We call this approach *LINK clustering*. Then we extract index terms, build the vector space and make clustering them. This traditional approach is called *VSM clustering*. Then we combine the two results into one to obtain new clustering of Web pages. Each group is not only characterized by some topic but also connected tightly with each other in a sense of contents.

EXAMPLE 31 *Let us show LINK clustering by an example. By interpreting nodes as Web pages and arcs as hyperlinks, we can represent Web pages by the graph, the number of references by in-degree and so on. Suppose there are 6 nodes* $a_1, ..., a_6$ *as in figure 2. Let* $From(a)$ *be a set of nodes staring from* a*, called a* set of out-arcs *of* a*, its cardinality is called* out-degree *of* a*. Similarly, the set of arcs with the ending node* b*,* $To(b)$ *is called a* set of in-arcs *of* b*, its cardinality is*

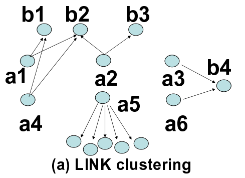


Fig. 2. LINK Clustering

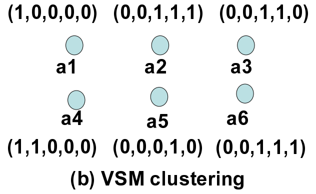


Fig. 3. VSM Clustering

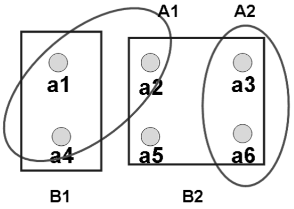


Fig. 4. Combination Clusters

called in-degree *of b. We apply complete link hierarchical clustering. This process is called LINK clustering. we get two LINK clusters A1 and A2 :A1=a1 , a2, a4 , A2 = a3,a6 As for a node a5, we consider a cluster of singleton and remove this.*

EXAMPLE 32 *VSM clustering is nothing but a clustering by document vectors. For example, given 6 Web pages* $P = \{a_1, \cdots, a_6\}$ *with the document vectors in figure 3, We apply complete link hierarchical clustering. This approach is called VSM clustering, and each cluster is called* VSM cluster. *We get 2 clusters* B_1, B_2*:* $B_1 = \{a_1, a_4\}$*,* $B_2 = \{a_2, a_3, a_5, a_6\}$ *Then let us* combine *the two results. In figure 4, two ovals represent LINK clusters* A_1, A_2 *while two rectangles describe VSM clusters* B_1, B_2*.*

EXAMPLE 33 *In example31 and 32, we obtain two combined clusters* $C_{11} = \{a_1, a_4\}$ *and* $C_{22}\{a_3, a_6\}$ *but we discard small clusters* $C_{12} = \{a_2\}$ *and* $C_{02} = \{a_5\}$ *so we got the final results of two groups* C_{11}, C_{22}*.*

3.2 Hierarchical Expression

Let us discuss technique of structuring each group of Web pages[13]. We assume sets of Web pages through combination clustering. Web pages are divided into small units, called *Semantic Textual Unit (STU)*, which are defined as a unit of meaningful contents to capture intended semantics of paragraphs or caption parts (in, say, `alt` tag) of Web pages. The similar approach is found in [3].

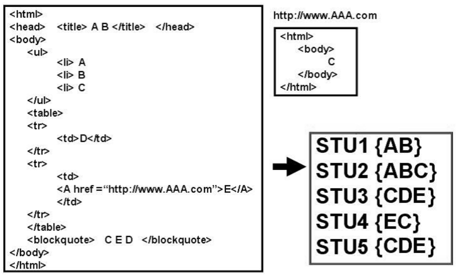
Well-formed Web pages contain strings parts and tag parts in a nested manner. Any strings quoted by a tag (i.e. `<tag>` ... `</tag>`) constitute an STU that carries meaningful unit of semantics intended by the `tag`. When the strings contain tag structures inside, the STU is called *nested*. When analyzing STU values, we extract STUs and their nested structure, called *STU (Nest)*. In this investigation, we examine `<UL> <OL>`, `<DL>`, `<TITLE>`, `<TABLE>` and `<BLOCKQUOTE>`.

One special tag `<A HREF>` describes *link* specification of the Web page, sometimes called a *context*. Here we replace the specification part by the contents of the linked Web pages. This is called *STU (Link)*.

All the words appeared in these STUs are represented by using vector space modeling. We should give *index terms* extracted from all the STUs. We take frequencies of the string expressions and reduce the dimensionality by using Zipf's law. *Similarity of two STUs* is now defined as the cosine value as usual.

We apply hierarchical algorithm to extract hierarchical relationship among Web pages. By examining distances between two clusters and combining the two of the minimum in a hierarchical manner, we get the hierarchical structures among the clusters (we do this *merge process* many time until we get to one big cluster). The results depend on the definition of distance (similarity) of clusters and the initial set of Web pages, and we should think about *how to construct* hierarchy. It is well-known that there are *single linkage method*, *complete linkage method* and *average linkage method* to define distance (similarity) of clusters, where we consider the distance of two clusters as the minimum, maximum and average distance of STUs between two clusters respectively. The first method is not suitable for our purpose since noisy STUs (accidental similarity) can't be avoided. By other two methods, we make up the hierarchy in this investigation. Moreover, to overview the whole relationship, we define *centroid* STU of each cluster as the nearest neighbor STU to the average values from the STU vectors, and build up the hierarchy consisting of the center STUs. To put labels to clusters, we introduce an option of centroids that are Representative sentence of the clusters. Because centroid comes from the definition of the similarity, our method could be seen as how to select the most important sentence as summarization.

EXAMPLE 34 *Let us explain our approach by example. In figure 5, given two Web pages and 5 words A,B,C,D and E within, we generate 5 STUs. Then we obtain hierarchical clustering by average linkage method in figure 6. A figure 6 contains the details of the result by average linkage method. As shown in figure 5, STU1 corresponds to a* `<TITLE>` *, STU2 to a* `<UL>` *and STU5 to a* `<BLOCKQUOTE>` *without any nesting. On the other hand, STU4 contains both a word E quoted by an* `<A>` *and a word C appeared in the linked Web page, STU3 contains STU3 as an inner structure. The similarity value is 1.000 which means it is generated by STU3 and STU5.*



```
<html>
<head>   <title> A B </title>   </head>
<body>
    <ul>
          <li> A
          <li> B
          <li> C
    </ul>
    <table>
    <tr>
          <td>D</td>
    </tr>
    <tr>
          <td>
          <A href =“http://www.AAA.com”>E</A>
          </td>
    </tr>
    </table>
    <blockquote>   C E D   </blockquote>
</body>
</html>
```

```
http://www.AAA.com
<html>
    <body>
          C
    </body>
</html>
```

STU1 {AB}
STU2 {ABC}
STU3 {CDE}
STU4 {EC}
STU5 {CDE}

Fig. 5. Two HTML files

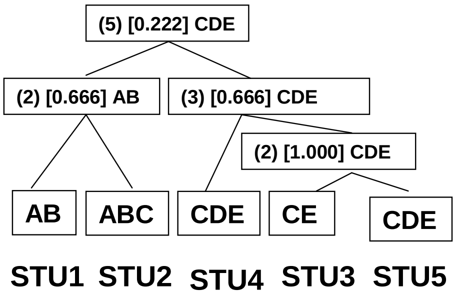


Fig. 6. ierarchical Expression

4 Evaluating Hierarchical Pages

When we summarize a set of Web pages, we wonder the answers are really correct or not. So is true for a case of machine translation. Even if we have *typical* answers by hand and examine the difference between summarization results and the typical answers, it is still hard for us to evaluate *how* well we get them[7]. This is because the judgment depends on subjectivity. One approach has been proposed where there should be agreement of the summarization results among the participants[11]. This means that summarization by hand is not always correct. In this section we discuss several evaluation methods of the automatic summarization[7], and put focus on the evaluation method of hierarchical topic detection. We propose three measures to evaluate hierarchical expression of Web pages, taht is, (1) Cluster readability, (2) Hierarchy readability and (3) Reading comprehension. We call the evaluation method as *CHR method* by taking the tree initial letters.

4.1 Evaluating Summarization

Generally, to evaluate summarization results, there have been proposed several measures so far[7]. The first point is *compression rate.* This means the ratio of length of sources and summarization. Clearly high compression rate doesn't contain all the contents of the source. The second is *Informativeness*, defined as measure of quantity to see how many contents of the source are preserved. The idea suggests that summarization should contain contents of source as a part. The compression rate and the informativeness are trade-off with each other.

Traditionally F-measure [3] has been proposed ans discussed. This measure corresponds to the evaluation of informativeness in both clustering and extracting. However we can't utilize F-measure to our case since it is hard to obtain all the answer in advance. Moreover, in the case of abstracting, it is difficult for us to evaluate *informativeness.*

Another approach comes from the viewpoints of *how many word are overlapped* by using *Dice coefficient* and the cosine similarity. Unfortunately, in this approach, we should resolve several issues of *synonym, homonym* and *construction information.* For example, we should think about thesaurus, Latent Semantic Indexing, syntactic analysis and conversation understanding of Natural Languages.

Let us put attention on the evaluation of readability and reading comprehension. *Readability* means the measure of "how well we can understand the summarization results". Mani[7] introduces some scoring techniques to dangling anaphor and a context. By *Reading comprehension* we examine the intelligibility of the reader who did read the summarization. Mani also introduces sophisticated methods[7].

4.2 Evaluating Hierarchical Topic Detection

What kinds of points should we pay attention on evaluation of the summarization for hierarchical expression ? There have been proposed some techniques for hierarchical topic detection, and let us discuss them as examples. The topic detection and tracking (TDT) task contains topic detection where news articles are organized into clusters which correspond to events (or topics)[1]. There exist two assumptions here about news articles. The first says that each news article describes only one event and is assigned to only one cluster. The second says

[3] F-measure combines *recall (r)* and *precision (p)* into one formula $F = \frac{2rp}{r+p}$ where *recall* means ratio of retrieved documents to all the answer documents while *precision* means correctness ratio.

that any hierarchical relationship among events should be ignored. As a result of the assumptions, any topic detection system assigns an article to among the non-hierarchical group. Obviously these assumptions don't reflect actual situation. A new Task Definition and Evaluation Plan of TDT 2004[2] has began for the purpose of *hierarchical topic detection* task. The evaluation method used for the old topic detection task is not suitable any more for this new task.

In TDT 2004, Allan has proposed a new evaluation method for hierarchical topic detection task[2]. This work has inspired many reserach activities. Trieschnigg examines three methods based on the method by Allan[14]. Especially they have revealed the fact that the detection cost isn't proportional to "power set" by putting focus on *false alarm* cost and *miss detection* cost, where a *power set* means to include all topics in a cluster. Fiscus and Doddington discusses a fact that miss detection costs more than false alarm[5]. To evaluate detection cost, Allan has introduced 4 combinations of some aspects defined by system output and clustering as in the table 1, where `in cluster` means an element is properly put into a cluster of interests (called "in *correct* cluster") and `relevant` means a system says the element is relevant to the cluster of interests. In the table, R_+, N_+, R_-, N_- mean the number of elements in each category respectively, given the number of all the elements n and the number of all the relevant elements r without any duplication. We define *miss detection ratio* P_{miss} of the cluster and *false alarm ratio* P_{fa} as follow :

Table 1. Four combinations

system output	relevant	non-relevant
in cluster	R_+	N_+
not in cluster	R_-	N_-
total	r	$n-r$

$$P_{miss} = \frac{R_-}{r} \tag{1}$$

$$P_{fa} = \frac{N_+}{n-r} \tag{2}$$

Finally we define *detection cost* as a linear sum of P_{miss} and P_{fa}.

$$C_{det} = C_{miss}P_{miss}P(target) + C_{fa}P_{fa}(1 - P(target)) \tag{3}$$

C_{miss} and C_{fa} are the costs of miss detection and false alarm respectively and *P(target)* is the prior probability to obtain the target.

Fiscus and Doddington discusses examines relationship between the costs and prior probabilities, that is, in TDT, *misses* should be penalized much more heavily than *false alarms* [5]. This is the reason they give $C_{miss} = 10$ and $C_{fa} = 1$. They assume a common value of *P(target)* for all topics based on corpus statistics; they give 0.02 as the constant. Eventually we define TDT cost function as:

$$C_{det} = 0.2P_{miss} + 0.98P_{fa} \tag{4}$$

Allan also have defined *travel cost* as the cost to find the most suitable cluster of each topic from the root node. This cost consists of a detection cost and depth to find the cluster. Here let us denote a depth from the root by D, then lst us define *travel cost* as follow:

$$dep = D/max(D) \tag{5}$$

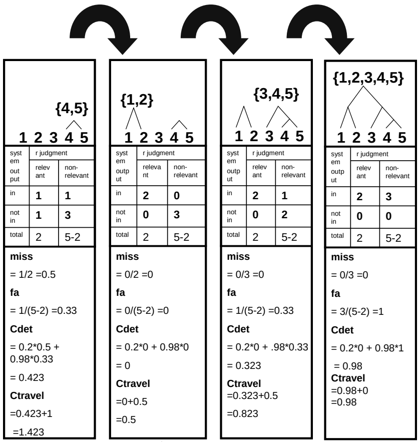
$$C_{travel} = C_{det} + dep \tag{6}$$

Finally, in Allan's approach, the minimal cost of a cluster is defined as the shortest path to this cluster from the structure's root cluster.

$$C_{minimal} = min(C_{travel}) \tag{7}$$

EXAMPLE 41 *Let us evaluate our example 34. We assume two clusters of {1,2},{3,4} are correct clusters. ans we compare a merged cluster {4,5} with a correct cluster {3,4}.*

R_- is assigned to {3}, and N_+ is assigned to {5}. The details of the calculation of each cost are shown in the figure7. Then we obtain the minimal cost as $C_{minimal} = 0.5$ in cluster {1,2}.



system output	r judgment	
	relevant	non-relevant
in	1	1
not in	1	3
total	2	5-2

miss
= 1/2 =0.5
fa
= 1/(5-2) =0.33
Cdet
= 0.2*0.5 + 0.98*0.33
= 0.423
Ctravel
=0.423+1
=1.423

system output	r judgment	
	relevant	non-relevant
in	2	0
not in	0	3
total	2	5-2

miss
= 0/2 =0
fa
= 0/(5-2) =0
Cdet
= 0.2*0 + 0.98*0
= 0
Ctravel
=0+0.5
=0.5

system output	r judgment	
	relevant	non-relevant
in	2	1
not in	0	2
total	2	5-2

miss
= 0/3 =0
fa
= 1/(5-2) =0.33
Cdet
= 0.2*0 + .98*0.33
= 0.323
Ctravel
=0.323+0.5
=0.823

system output	r judgment	
	relevant	non-relevant
in	2	3
not in	0	0
total	2	5-2

miss
= 0/3 =0
fa
= 3/(5-2) =1
Cdet
= 0.2*0 + 0.98*1
= 0.98
Ctravel
=0.98+0
=0.98

Fig. 7. Allan's Calculation

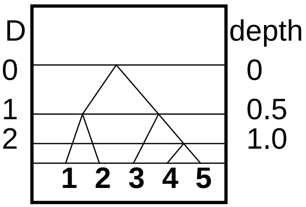


Fig. 8. Depth Calculation

In Allan's approach, we see several problems. One problem is that *correct clusters* are not realy *correct* hierarchical clusters. For example, in a cluster {1,2,3,4,5} after merged in the example above, a cluster {1,2} is computed as relevant elements while a cluster {3,4,5} is computed as non-relevant elements, which is not suitable since correct clusters of non-hierarchical clustering are exclusive. It should be problematic to utilize correct non-hierarchical clusters for hierarchical clustering.

Another problem comes from *Power set*. We have $C_{miss} = 0$ and $C_{fa} = 1$ when combining {1,2} and {3,4,5}. Let us note that *Power set* can't be evaluated in terms of false alarm cost. We obtain $C_{fa} = N_+/(270000 - r)$ if we have huge number of elements. It is noted that, when the number r of the cluster element is small (for instance, if a cluster has a deep path from root node), *Power set* can't be evaluated as the example.

4.3 CHR Method

We propose a novel evaluation method using both correlation and hierarchy aspects. Our basic idea is that hierarchical expression describes the relations among the cluster by using correlation. We propose 3 kinds of evaluation methods for *cluster readability, hierarchy readability* and *reading comprehension.*

Let us discuss how to evaluate the readability of cluster. Cluster granularity means "the roughness of the cluster" but not "size of clusters". We say the granularity is *compact* if cluster elements are similar to each other. Otherwise we say the granularity is *coarse.* For instance, a cluster of 26 elements $A - Z$ and a cluster of 26 elements containing only A have the same size of cluster but the latter cluster is easier to grasp contents. We believe strong relationship between the readability of cluster and the granularity. Therefore, we evaluate the granularity by means of Allan's *detection cost.*

Allan has proposed *detection cost* consisting of two measures C_{fa} and C_{miss}. By C_{fa} they evaluate how dissimilar elements in a cluster are with each other, and by C_{miss} they evaluate how many similar elements in a cluster they miss.

According to the ideas, we introduce a notion of *cluster granularity* consisting of *inner* and *outer* granularities. Correponded to C_{fa} by Allan, we define inner-granularity by C_{in} as follow:

$$C_{in} = 1 - \frac{2}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i}^{m} sim(x_i, x_j) \tag{8}$$

where m means the number of elements in a cluster, $x_k(k : 1 \ldots m)$ means an element in a cluster, and $sim(x_i, x_j)$ means similarity between the elements. This cost means how coarse the cluster elements are related.

Similarly we define outer-granularity by C_{out} as follows:

$$C_{out} = \sum_{i=1}^{s} \sum_{j=i}^{s} sim(Cl_i, Cl_j) \tag{9}$$

where s means the number of clusters, $Cl_r(r : 1 \ldots s)$ means a cluster, and similarity between Cl_i and Cl_j clusters is denoted by $sim(Cl_i, Cl_j)$ based on average-linkage method. This cost represents how similar with each other the clusters are.

Then we restate C_{det} by the two costs, where C_{det} means the *readability of cluster.*

$$C_{det} = C_{in} + C_{out} \tag{10}$$

Note that the cost means how compact a cluster is and how similar the cluster is to other ones.

Let us introduce *Cophenetic coefficient correlation* for the purpose of evaluation of hierarchical relationship. Let us show how we evaluate two hierarchical clusters by similarity matrix.

Assume that we have similarity 0.0 between the element 1 and 3 as shown in the similarity matrix. When we combine {1,2} and {3} as in figure 9, we get the

	1	2	3
1	0.0	1.0	0.0
2	0.0	0.0	0.8
3	0.0	0.0	0.0

Table 2. Similarity Matrix

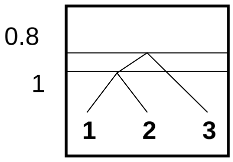


Fig. 9. Hierarchical Structure 1

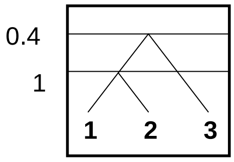


Fig. 10. Hierarchical Structure 2

similarity 0.8 between 1 and 3 based on single linkage. In a figure 10, we get the similarity 0.4 by average linkage. In any cases, when two clusters are combined, similarity should be changed, so that we can't preserve the original similarity matrix, and the value depends on what kind of linkage we have.

The idea of *cophenetic coefficient* correlation provides us both to maintain similarity matrix and to capture current situation of hierarchical clustering. We can describe hierarchy by *cophenetic matrix.* Elements of similarity matrix x and cophenetic matrix y are obtained by means of *pearson product-moment correlation coefficient*, defined as follow:

$$r_{x,y} = \frac{\sum xy - (1/n)(\sum x)(\sum y)}{\sqrt{\{\sum x^2 - (1/n)(\sum x)^2\}\{\sum y^2 - (1/n)(\sum y)^2\}}} \tag{11}$$

This is called *cophenetic coefficient correlation.* If the value is close to 1, there exists positive correlation, and if it is close to -1, there exists negative correlation.

As described, we evaluate *reading comprehension* by the length of a path from the root to the most suitable clusters. Hierarchical expression contains informativeness latently as much as sources, because the expression assigns all STU's to leaf nodes. Remember the *reading comprehension* increases whenever clusters are shallow and informative.

We can evaluate the situation by two costs C_{travel} and $C_{minimal}$. Let D be a depth from the root and define the costs as follows.

$$dep = D/max(D) \tag{12}$$

$$C_{travel} = C_{det} + dep \tag{13}$$

$$C_{minimal} = min(C_{travel}) \tag{14}$$

EXAMPLE 42 *Let us apply CHR method to our example. The table 3 contains similarity matrix of example 34. Figure12 illustrates how* cluster readability *and* reading comprehension *are calculated.*

Table 3. Our Similarity Matrix

	1	2	3	4	5
1	0.0	0.66	0.0	0.0	0.0
2	0.0	0.0	0.22	0.22	0.22
3	0.0	0.0	0.0	0.66	0.66
4	0.0	0.0	0.0	0.0	1.0
5	0.0	0.0	0.0	0.0	0.0

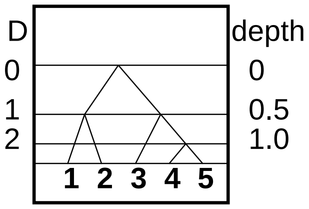
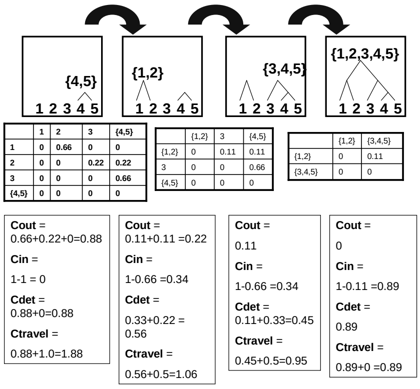


Fig. 11. Detail of Depth Calculation



	1	2	3	{4,5}
1	0	0.66	0	0
2	0	0	0.22	0.22
3	0	0	0	0.66
{4,5}	0	0	0	0

	{1,2}	3	{4,5}
{1,2}	0	0.11	0.11
3	0	0	0.66
{4,5}	0	0	0

	{1,2}	{3,4,5}
{1,2}	0	0.11
{3,4,5}	0	0

Cout = 0.66+0.22+0=0.88
Cin = 1-1 = 0
Cdet = 0.88+0=0.88
Ctravel = 0.88+1.0=1.88

Cout = 0.11+0.11 =0.22
Cin = 1-0.66 =0.34
Cdet = 0.33+0.22 = 0.56
Ctravel = 0.56+0.5=1.06

Cout = 0.11
Cin = 1-0.66 =0.34
Cdet = 0.11+0.33=0.45
Ctravel = 0.45+0.5=0.95

Cout = 0
Cin = 1-0.11 =0.89
Cdet = 0.89
Ctravel = 0.89+0 =0.89

Fig. 12. Detail of Cost Calculation

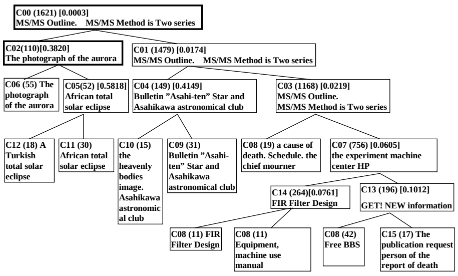

C00 (1621) [0.0003] MS/MS Outline. MS/MS Method is Two series
C02(110)[0.3820] The photograph of the aurora
C01 (1479) [0.0174] MS/MS Outline. MS/MS Method is Two series
C06 (55) The photograph of the aurora
C05(52) [0.5818] African total solar eclipse
C04 (149) [0.4149] Bulletin "Asahi-ten" Star and Asahikawa astronomical club
C03 (1168) [0.0219] MS/MS Outline. MS/MS Method is Two series
C12 (18) A Turkish total solar eclipse
C11 (30) African total solar eclipse
C10 (15) the heavenly bodies image. Asahikawa astronomical club
C09 (31) Bulletin "Asahi-ten" Star and Asahikawa astronomical club
C08 (19) a cause of death. Schedule. the chief mourner
C07 (756) [0.0605] the experiment machine center HP
C14 (264)[0.0761] FIR Filter Design
C13 (196) [0.1012] GET! NEW information
C08 (11) FIR Filter Design
C08 (11) Equipment, machine use manual
C08 (42) Free BBS
C15 (17) The publication request person of the report of death

Fig. 13. Hierarchical Summarization

5 Experimental Results

In this section we show some experimental results to see how effective our approach works. Here we examine a test collection of Web pages, called NTCIR-3, provided by NII Japan. Here we discuss the two experiments to examine our algorithms to one of the clusters. In the first experiment, we compare our CHR method to Allan's method and examine the evaluation results. In the second experiment, we compare several STU to see how well they play important roles for hierarchical summarization.

5.1 Preliminaries

Before describing our results, let us review our previous work quickly. In this experiment, `NTCIR-3` collection contains many `HTML` and textual data in Japanese internet (i.e., in `.jp` domain). We have selected 9,929 pages dated September 29 to October 5 in 2001 to be examined. We have applied the combination clustering to the data and and we got 6 meaningful clusters.

By putting our focus on a Cluster, let us discuss Hierarchical expression. A figure 13 contains result of the hierarchical summarization generated by average linkage where the digits in `(..)` and `[..]` show the number of STU elements and the similarity at cluster merging.

Let us examine each summarization result and corresponding cluster. There is leaf node STU which correspond to Web page contained in this group. In fact, "experiment equipments" appeared in a cluster 1 correspond to clusters C00/ C01/ C03/ C07/ C14/ C17/ C18 while "a cause of death" corresponds to C08, "free BBS" to C16. There exists an STU about "Asahikawa astronomical club", and we can see *topic drifting* because many STUs about "Asahikawa university" arise in this cluster.

A parent node describes the more abstract contents compared to the child nodes. For example, the number of African total solar eclipse cluster (C12) elements is thirty pages. Turkish total solar eclipse cluster (C11) contains 18 pages. When combining them (C05), we get African total solar eclipse as a label. Then a bigger cluster becomes an abstraction of those two. This means that the higher nodes contain, the more STUs we have and the more frequent topics we see.

It is common that the higher level STUs describe, the more general contents we have. In fact, the highest level STUs are Asahikawa university and digital filter lab which are the dominant pages.

To put labels to clusters, we introduce a centroid that can be seen as a representative sentence of the clusters. Because a centroid comes from the definition of the frequency, our method provides us to extract the most important sentence hierarchically. Also the centroid approach is consistent with hierarchical structure of clusters.

5.2 Experiment 1

As shown in the previous subsection, we examine 1621 STUs from 35 clusters. In this experiment, let us compare our CHR method with Allan's. We calculate costs defined by CHR method as well as Allan method, and compare C_{in} with C_{fa}, and C_{out} with C_{miss}.

When two clusters $C1, C2$ are combined into one, say Cm, let us discuss how many false alarm and errorneous element we have with respect to Allan's method, as shown in table4. In the table, when the two clusters C, G are put together, we see C_{miss} decreases. Generally Allan's C_{miss} never increases when we combine clusters, because every article carries only one event (topic) and the topic of the lowest C_{miss} is chosen. For example, C_{miss} cannot increase at step 3 in the table 4. However, when a cluster contains more than one topics, we should calculate C_{miss} over all topics, and the table 5 shows C_{out} increases according to the CHR method.

Allan's cost C_{fa} depends on the number of cluster elements and the number of all the elements. When clusters are combined, all the elements except elements in a certain topic are classified as N_+. This means C_{fa} increases very much when we get a cluster of many elements. For example, in the table 4, C_{fa} at $step1$ and $step2$ are small, but C_{fa} increases in the step 3 drastically. On the other hand, CHR method allows us to calculate granularity and the costs come based on similarity between topics in a cluster but not heavily on the number of elements. In our case, C_{in} doesn't increase at step 3 drastically as in the table 5.

Table 4. Allan's *miss* and *fa*

step	C1	C2	Cm	C_{miss}	C_{fa}
1	A	B	C	0.517241	0.0712743
2	E	F	G	0.873418	0.0155642
3	C	G	H	0.107759	0.419726

Table 5. CHR's *miss* and *fa*

step	C1	C2	Cm	C_{out}	C_{in}
1	A	B	C	0.004692449	0.9083263
2	E	F	G	0.001672507	0.9379648
3	C	G	H	0.002181099	0.9438765

We show the average of C_{fa} and C_{miss} in each depth in figure 14, to compare C_{out} with C_{miss}, and C_{in} with C_{out}. The depth at which two costs are the smallest (by CHR method) corresponds to the one by the Allan's method as shown in figure15. However, in figure14, CHR method doesn't correspond to Allan's method at all. Allan gives some weight with some sort of approximation to come close to ideal detection cost. By CHR method we get similar results to Allan, but with no weight and no correct clusters given in advance.

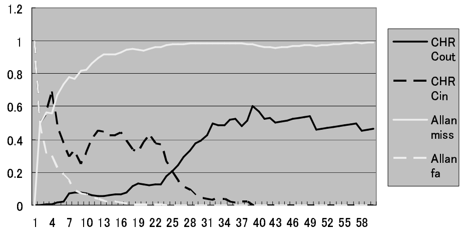


Fig. 14. *miss* Cost and *fa* Cost

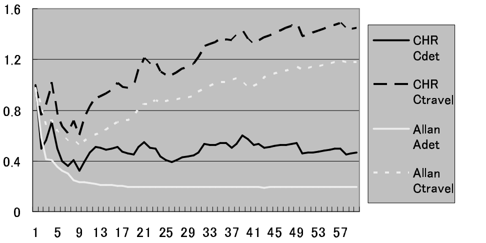


Fig. 15. Detection Cost and Travel Cost

5.3 Experiment 2

Our next experiment concerns on a fact that links and nesting are really effective to obtain better summarization. We examine how *STU(Nest)* and *STU(Link)* are useful. In this experiment, we generate STUs by means of 4 types of constructs: STU, STU(Nest), STU(Link), and STU(Nest+Link), which mean that we have STU without any nesting, with nesting, with link specification but not nesting, and STU with nesting and link. We obtain $\hat{y}$ in such a way that:

$$\hat{y} = argmin_{y \in Y}(Cdet(y)) \tag{15}$$

This equation means that $\hat{y}$ is predominant in C_{det}. $C_{travel}(u)$ and $cophenetic(v)$ are defined similarly.

$$\hat{u} = argmin_{u \in Y}(Ctravel(u)) \tag{16}$$

$$\hat{v} = argmax_{v \in Y}(cophenetic(v)) \tag{17}$$

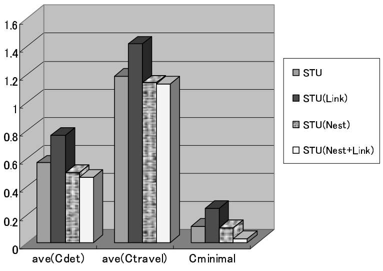


Fig. 16. Calculating *argmin*

Fig. 17. Cophenetic Correlation Coefficient

	cophenetic correlation coefficient
STU	0.678701
STU (Link)	0.775679
STU (Nest)	0.667447
STU (Nest+Link)	0.717835

We examine the three argmin/argmax values by means of one of the 4 methods, the results are shown in the figure16.

We see *STU(Nest)* wins the best result in figure16. We found some STUs appear also in STU(nest), in this case, we must have compact granularity. Let us illustrate the results of cophenetic correlation coefficient in table 17. Here *STU(Link)* shows the best results, this means referenced pages tells us main topics generally. When we replace links by representative contents, we should have heavy change of centroids. This means *STU(Link)* is one of the key factors which improves cophenetic coefficient correlation.

By looking at these results, we conclude two aspects as follows. The first is that *STU(Nest)* plays important role on both detection cost and travel cost. Second is that *STU(Link)* improves cophenetic coefficient correlation essentially. Because *STU(Nest+Link)* carries these aspects, we can say CHR method is excellent.

6 Conclusion

In this investigation, we have proposed a new technique to summarize contents of Web pages in a form of hierarchical expression. And then, we have proposed a novel evaluation method for the expression. Experiments show the good results and we can say the approach is promising.

Our approach starts with decomposition of pages into STUs and application of hierarchical clustering. An automated procedure can be implemented easily.

References

1. Allan, J., Carbonell, J., Doddington, G., Yamron, J. and Yang, Y. : Topic detection and tracking pilot study: Final report, the DARPA Broadcast News Transcription and Understanding Workshop, 1998
2. Allan, J., Feng, A. and Bolivar, A.: Flexible Intrinsic Evaluation of Hierarchical Clustering for TDT, the 12th international Conference on Information and Knowledge Management (CIKM), 2003, pp.263 - 270
3. Buyukkokten, O., Garcia-Molina, H. and Paepcke, A.: Seeing the Whole in Parts:Text Summarization for Web Browsing on Handheld Devices, In Proceedings International WWW Conference, 2001
4. Chakrabat,S.: Mining the Web, Morgan Kaufmann, 2003
5. Fiscus, J,G. and Doddington, G.R. : Topic detection and tracking evaluation overview, Event-based Information Organization, 2002, pp. 17-31
6. Jain, A.K., Murty, M. N. and Flynn, P. J.: Data clustering: a review, ACM Computing Surveys, 31-3, 1999, pp.264-323
7. Mani,I. : Automatic Summarization, John Benjamins, USA, 2001
8. Mori, M., Miura, T and Shioya, I.: Extracting Events From Web Pages, International Conference on Advances in Intelligent Systems - Theory and Applications
9. Mori, M., Miura, T and Shioya, I.: Abstracting Temporal Clusters, Internet Technologies and Applications (ITA), 2005
10. Yukio Ohsawa, Nels E. Benson and Masahiko Yachida: KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor Proc. Advanced Digital Library Conference (IEEE ADL'98), pp.12-18, 1998
11. Radev, D., Jing, H. and M. Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, Information Processing and Management, 2004, pp.919-938
12. Takahashi, K., Miura, T and Shioya, I.: Combination Clustering for Web Correlation, IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), 2005, pp.434-437
13. Takahashi, K., Miura, T and Shioya, I.: Hierarchical Summarization of Web Pages, IADIS Applied Computing, 2006, pp.612-617
14. Trieschnigg, D. and Kraaij, W.: Scalable Hieararchical Topic Detection, SIGIR, 2005
15. Zamir, O. and Etzioni, O.: Web Document Clustering: A Feasibility Demonstration, the 21st International ACM SIGIR Conference, 1998, pp.46-54