User-Focused Multi-Document Summarization with Paragraph Clustering and Sentence-Type Filtering

Yohei Seki†,‡ Koji Eguchi‡, † Noriko Kando‡, †
†Department of Informatics, The Graduate University for Advanced Studies (Sokendai)
‡National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
seki@grad.nii.ac.jp eguchi@nii.ac.jp kando@nii.ac.jp

Abstract

Applying document clustering techniques to multi-document summarization is a challenging problem, mostly because of the redundancy that exists in multiple sources. We compare several document clustering techniques for multi-document summarization in the NTCIR-4 TSC test collection. We conducted an experiment to evaluate the effectiveness of reducing redundancy in the production of summaries. From the results, we draw conclusions regarding the nature of the multi-document summarization with respect to redundancy reduction strategies.

Keywords: *Multi-document Summarization, User-focused Summarization, Redundancy Elimination, Document Clustering.*

1 Introduction

The goal of multi-document summarization (MDS) is usually defined as *to extract content from a collection of related documents and present the most important content sensitive to the user's needs. What is required is that the similarities and differences be taken into account, along with redundancy in the course of producing the summary* [10]. On the other hand, document clustering techniques partition a set of objects into clusters. Van Rijsbergen's *cluster hypothesis* suggests that *closely associated documents tend to be relevant to the same request* [20]. The cluster hypothesis suggests that we could cluster the documents about a given topic and retrieve them all in response to a query. Therefore, these techniques are closely related to the goal of multi-document summarization.

Many clustering-based multi-document summarization frameworks [18, 19, 12, 14, 2, 6, 1] have been proposed, as shown in Table 1. Research on sentence redundancy was proposed in the TREC 2003 Novelty Track [21]. These methods have four principal aspects: (1) clustering algorithms, (2) cluster units, (3) sentence extraction strategy, and (4) cluster size.

This paper is organized as follows. The next section provides a brief overview of our summarization algorithm. In Section 3, we will show you our results in NTCIR-4 TSC Evaluation. We then present post-submission analysis mainly for clustering techniques to produce better summaries. Finally, we present our study of sentence-type filtering approach to improve the responsiveness and show our conclusions.

2 Clustering-based Summarization Approach for Reducing Redundancy

The source documents could be clustered by single document summary, sentence or paragraph units. We clustered source documents by not by sentence units but by paragraph units for the following reasons: (1) it allowed real-time interactivity, and (2) because of the sparseness of sentence vectors. In addition, we did not cluster source documents by document units because source document sizes (from 5 to 19 documents) were too small compared to summary sizes (from 4 to 32 sentences).

We now describe our clustering-based multi-document summarization algorithm. This algorithm was constructed as two main stages: paragraph clustering and sentence extraction.

Algorithm 1 Our Clustering-based Multi-document Summarization

▷ *1. Paragraph Clustering Stage*

Source documents were segmented to paragraph units, then features with term frequencies were computed for each paragraph.

Paragraph units were clustered with term-frequency similarity. A clustering algorithm (complete link, group average, or Ward's method[4]) was selected and applied. Cluster sizes were changed based on the number of extracted sentences.

▷ *2. Sentence Extraction Stage*

The feature vectors for each cluster were computed with term frequencies and inverse cluster frequencies.

if questions focusing on a summary were given

Table 1. Comparison of Clustering-based MDS Frameworks

Author [References]	Algorithm	Unit	Similarity (Distance)	Feature	Extract Strategy
Stein et al.[18, 19]	Complete link	Single Document Summary	Dice coefficient	TF	Sentences similar to the cluster centroid.
Radev et al.[14]	Single Path	Document	Cosine coefficient	TF*IDF	Sentence weights with centroid, position, and MMR.
Boros et al.[2]	k-means	Sentence	Cosine coefficient	TF	Sentences similar to the cluster centroid.
Hatzvassiloglou et al.[6]	Exchange	Paragraph	Overlapping Feature	log-linear model.	One sentence from each cluster with heuristics and covering words.
Moens et al.[12, 1]	Covering k-medoid	Single Document Summary/Paragraph	Cosine coefficient	TF	Sentences closer to medoid of their cluster.
Zhang et al.[21]	Unknown	Sentence	Unknown	subtopic, date/opinion, sentence vector	Sentences similar to a query.

clusters were ordered by the similarity between content words in the questions and the cluster feature vectors.

else

we computed the total term frequencies of all documents and ordered clusters based on similarities between total TF and cluster feature vectors.

end

Sentences in each cluster were weighted based on question words, heading words in the cluster , and TF values in the cluster.

One or two sentences were extracted from each cluster in cluster order to reach the maximum allowed number of characters or sentences.

3 NTCIR-4 TSC Official Evaluation

In NTCIR-4 TSC, four evaluations were applied to 9 participants' systems [7]. Our system ID was F0301.

3.1 Extract Evaluation

The extraction evaluation results was shown in Table 2. Our system's official submission result that corresponded to the official abstract result was F0301(a) and the clustering algorithm used for it was based on the "group average" method. F0301(b) was a second official submission result that did not correspond to abstract results and used a clustering algorithm based on the "Ward's method". Of the 9 teams, the "coverage" of F0301(a) averaged over short and long summaries was ranked second, and its "precision" was ranked third. You could refer the definition of "coverage" and "precision" to [7]. Our system was effective for redundancy elimination because whereas the "coverage" measure counted the overlapping elements, the "precision" measure did not.

3.2 Abstract Responsiveness Evaluation

Abstract responsiveness evaluation was a simulated extrinsic evaluation. The results are shown in Table 3. Our system ID was again F0301. Of the 9 teams, our system averaged over short and long summaries was ranked second for both exact matches and edit distances.

Table 2. Extract Evaluation

	Short		Long	
ID	Cov.	Prec.	Cov.	Prec.
F0301(a)	**0.315**	**0.494**	**0.355**	**0.554**
F0301(b)	**0.372**	**0.591**	**0.363**	**0.587**
F0303(a)	0.222	0.314	0.313	0.432
F0303(b)	0.293	0.378	0.295	0.416
F0304	0.328	0.496	0.327	0.535
F0306	0.283	0.406	0.341	0.528
F0307	0.329	0.567	0.391	0.680
F0308	-	-	-	-
F0309	0.308	0.505	0.339	0.585
F0310	0.181	0.275	0.218	0.421
F0311	0.251	0.476	0.247	0.547
LEAD	0.212	0.426	0.259	0.539
HUMAN	-	-	-	-

Cov. = coverage
Prec. = precision

3.3 Abstract Content Evaluation

We show manual content evaluation for abstracts in Table 4. Of the 9 teams, our system was ranked fifth averaged over short and long summaries. We re-examined our submission results and found bugs that caused over-size summaries for nine short summaries and 15 long summaries. In addition, our summary results were not arranged correctly. We revised the system, so that bugs were fixed and the sentences in the summaries were ordered chronologically by news sources.

3.4 Quality Questions

We show the evaluation results for Quality Questions in Table 5. Sixteen quality questions were evaluated manually for readability tests. Although the evaluation results were not so good in total because we did

Table 5. Evaluation for the Quality Question

Topic	Q0	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
310S	0	1	0	0	1	0	0	0	-1	2	3	0	0	0	0	0
310L	0	0	0	0	2	3	0	0	-1	0	0	0	0	0	0	0
320S	0	1	4	1	1	1	3	3	-1	1	0	1	0	0	0	1
320L	0	6	2	1	9	1	7	2	-1	1	5	0	0	0	0	1
340S	0	2	1	0	2	0	1	0	1	0	0	0	0	0	0	0
340L	0	3	0	1	1	1	0	0	-1	1	2	0	0	0	0	0
350S	1	2	0	0	5	0	1	3	1	1	9	0	0	0	1	0
350L	0	5	1	1	11	2	1	1	1	1	15	1	0	0	0	0
360S	0	2	0	0	0	0	0	0	0	1	0	0	0	0	1	0
360L	0	3	0	0	1	3	0	0	0	0	0	0	0	0	1	2
370S	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0
370L	0	1	3	0	1	3	2	1	-1	1	0	0	0	0	0	0
380S	1	2	2	2	3	1	0	0	-1	1	0	0	0	0	0	0
380L	0	0	1	1	2	2	0	0	-1	1	2	0	0	0	0	0
400S	0	0	0	0	1	5	0	0	-1	3	0	1	0	0	0	0
400L	0	1	0	0	5	2	1	0	-1	2	4	0	0	0	0	0
410S	0	1	0	0	4	1	4	3	-1	0	4	0	0	0	0	0
410L	0	8	3	1	4	0	3	10	-1	1	1	3	0	0	0	0
420S	1	1	0	0	3	0	0	3	-1	0	5	0	0	0	0	1
420L	1	0	0	0	2	1	1	0	-1	0	0	0	0	0	0	0
440S	0	3	2	1	1	0	1	0	-1	0	0	0	0	0	0	0
440L	0	0	1	1	5	0	0	0	-1	3	4	1	0	0	0	1
450S	0	0	0	0	0	2	0	0	0	1	1	0	0	0	0	0
450L	0	1	0	0	2	5	0	1	0	0	1	0	0	0	0	1
460S	0	0	0	0	0	0	0	0	-1	0	1	0	0	0	0	0
460L	0	0	2	0	1	0	1	0	-1	1	0	0	0	0	0	0
470S	1	1	0	1	1	1	1	0	-1	0	0	0	0	0	0	0
470L	0	3	0	0	10	0	1	3	-1	3	2	1	0	0	0	0
480S	1	0	0	1	1	0	0	0	-1	0	1	0	0	0	0	1
480L	1	0	0	0	5	1	0	1	-1	0	4	2	0	0	0	0
500S	0	0	0	0	1	1	0	0	-1	0	0	0	0	0	0	0
500L	0	0	1	0	5	2	0	0	-1	1	0	0	0	0	0	0
510S	1	1	1	0	5	1	2	0	-1	0	0	0	0	0	1	0
510L	1	0	3	0	7	0	5	6	-1	0	4	0	0	0	1	0
520S	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
520L	0	1	0	0	6	0	2	3	-1	0	2	0	0	0	0	0
530S	0	3	0	0	5	0	1	4	-1	1	0	0	0	0	0	0
530L	0	3	3	1	13	3	1	10	-1	1	0	0	0	0	0	0
540S	0	2	1	0	4	2	1	0	-1	0	2	0	0	0	0	0
540L	0	2	1	0	13	2	0	0	-1	0	0	0	0	0	0	1
550S	1	0	0	0	2	3	2	0	-1	1	4	1	0	0	0	0
550L	1	0	1	0	8	3	1	0	-1	2	3	2	0	0	0	2
560S	0	2	0	3	8	2	0	4	-1	1	6	0	0	0	0	1
560L	0	0	0	0	11	2	3	4	-1	0	13	0	0	0	0	0
570S	0	0	2	1	4	0	1	1	-1	0	1	0	0	0	0	0
570L	0	0	1	0	11	2	0	6	-1	0	4	0	0	0	0	1
580S	1	1	1	0	1	0	0	1	1	0	0	0	0	0	0	0
580L	1	1	6	2	3	5	0	1	0	0	0	2	0	0	0	0
590S	1	0	0	0	1	0	0	2	0	0	4	0	0	0	0	0
590L	1	1	0	3	5	0	0	3	-1	0	8	0	0	0	0	0
600S	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	0
600L	0	1	0	0	2	3	0	0	-1	1	0	0	0	0	0	0
610S	0	1	2	0	2	1	0	1	-1	0	0	0	0	0	0	0
610L	0	1	4	1	7	1	1	0	-1	0	1	0	0	0	0	0
630S	0	0	2	0	6	4	0	0	-1	1	4	0	0	0	0	0
630L	0	0	0	0	6	10	0	0	-1	1	0	0	0	0	0	0
640S	0	3	0	0	4	0	0	0	-1	0	0	0	0	0	0	0
640L	0	5	1	0	5	3	0	0	-1	1	0	0	0	0	0	0
650S	0	0	0	0	2	1	1	3	-1	0	8	0	0	0	0	0
650L	0	2	0	0	3	3	0	5	-1	0	14	3	0	0	0	0

Table 3. Responsiveness Evaluation

	SHORT		LONG	
ID	exact	edit	exact	edit
F0301	**0.394**	**0.677**	**0.399**	**0.706**
F0303	0.257	0.556	0.266	0.602
F0304	0.367	0.653	0.356	0.677
F0306	0.342	0.614	0.327	0.630
F0307	0.439	0.710	0.442	0.751
F0308	0.321	0.601	0.313	0.611
F0309	0.390	0.684	0.356	0.633
F0310	0.133	0.427	0.201	0.549
F0311	0.304	0.579	0.308	0.628
LEAD	0.300	0.589	0.275	0.602
HUMAN	0.461	0.716	0.426	0.721

exact = exact match
edit = edit distance

Table 4. Abstract Content Evaluation

ID	Short	Long
F0301	**0.228**	**0.214**
F0303	0.188	0.240
F0304	0.247	0.258
F0306	0.230	0.248
F0307	0.291	0.323
F0308	0.222	0.210
F0309	0.207	0.247
F0310	0.131	0.233
F0311	0.197	0.221
LEAD	0.160	0.159
HUMAN	0.385	0.402

not put much effort into readability in our official results, the editing of conjunctions (Q6) was evaluated third because our system omitted them.

4 Post-submission Analysis

Our summarization method was based on a two-stage process: (1) paragraph clustering by topic, and (2) sentence extraction from clusters. In this section, we describe post-submission analyses from these two aspects.

4.1 Feature Vector and Distance Measure for Paragraph Clustering

To make correct clusters, feature vectors and distance measures were important because they changed the cluster structure drastically. We changed two feature vectors, raw term frequency and normalized term frequency, in each document. We also changed two distance measures: cosine similarity-based distance and euclidean distance. The results are shown in Table 6. We found raw term frequency and euclidean distance were effective for clustering to produce better summaries.Vector-length normalization typically does not work well for short documents [15]. Paragraphs are shorter units than documents, so this finding could also be applied to our paragraph clustering methods.

4.2 Clustering Methods

We compared several hierarchical clustering algorithms, as shown in Table 7: the complete link method, the group average method, and the Ward's method. We also compared the relationship between the number of extracting sentences and the cluster size. Our experiments showed that the Ward's method worked better than the other two methods. In addition, small cluster numbers (1.5 $\times$ sentence numbers to extract short summaries, and 1 $\times$ sentence numbers to extract long summaries) for the Ward's method worked best. The Ward's method has also been reported [3] to perform well compared to several agglomerative clustering methods, so these results matched our intuition.

In addition, we compared numbers of extracted sentences. We extracted one or two sentences from each cluster by finding query words up to the limit of sentence numbers. The result showed that the one-sentence extraction strategy performed better than the two-sentences extraction strategy.

4.3 Sentence Extraction Clues

We surveyed sentence extraction clues. Our system [16] in the previous NTCIR-3 TSC [13, 9, 5] used three types of clues to produce summaries: relative

Table 6. Feature Vector and Distance Measure for Paragraph Clustering

Algorithm		No Clustering	Complete Link		
Distance		-	Euclidean	1-cosθ	Euclidean
Unit		Paragraph			
Feature		Term frequency vectors for nouns and unknown words			Normalized TF
Cluster Number for Long Summaries (× Extracting Sentence Number)		-	× 1		
Cluster Number for Short Summaries (× Extracting Sentence Number)		-	× 1.5		
Extract one sentence from each cluster	Coverage	0.339	**0.358**	0.307	0.317
	Precision	0.614	**0.522**	0.398	0.429
	Redundancy (=Precision-Coverage)	0.275	**0.164**	0.091	0.112
Extract two sentences from each cluster	Coverage	0.319	0.327	0.322	0.325
	Precision	0.601	0.578	0.513	0.525
	Redundancy	0.282	0.251	0.191	0.200

Table 7. Coverage and Precision Change for Algorithms and Cluster Sizes

Algorithm		Complete Link			Group Average			Ward's Method			No Clustering
Distance		Euclidean									-
Unit		Paragraph									
Feature		Term frequency vectors for nouns and unknown words									
Cluster Number for Long Summaries (× Number of Sentences Extracted)		× 1	× 1.5	× 2	× 1	× 1.5	× 2	× 1	× 1.5	× 2	-
Cluster Number for Short Summaries (× Number of Sentences Extracted)		× 1.5	× 2	× 2.5	× 1.5	× 2	× 2.5	× 1.5	× 2	× 2.5	-
Extract one sentence from each cluster	Coverage	0.358	0.354	0.355	0.314	0.338	0.359	**0.364**	0.357	0.353	0.339
	Precision	0.522	0.544	0.567	0.499	0.543	0.579	**0.518**	0.543	0.565	0.614
	Redundancy (=Precision-Coverage)	0.164	0.190	0.212	0.185	0.205	0.220	**0.154**	0.161	0.212	0.275
Extract two sentences from each cluster	Coverage	0.327	0.334	0.324	0.317	0.321	0.317	0.334	0.323	0.315	0.319
	Precision	0.578	0.591	0.596	0.557	0.578	0.587	0.593	0.584	0.598	0.601
	Redundancy	0.251	0.257	0.272	0.240	0.257	0.270	0.259	0.261	0.283	0.282

Table 8. Sentence Extraction Clues

Algorithm		No Clustering			
Unit		Paragraph			
Clues	Heading	Yes	Yes	No	Yes
	Term Frequency	Yes	Yes	Yes	No
	Position	No	Yes	No	No
Extract one sentence from each cluster	Coverage	0.339	0.322	0.338	0.315
	Precision	0.614	0.606	0.613	0.623
	Redundancy (=Precision-Coverage)	0.275	0.284	0.275	0.308

position, heading words, and term frequencies in documents. We surveyed the coverage and precision of summaries using combinations of these clues with no clustering algorithm, as shown in Table 8. We found that "relative position" did not contribute to producing better summaries, but "term frequencies" and "heading words" did contribute. We used a strategy based on these discoveries in the official submissions. In addition, we used query words from the questions given by the NTCIR-4 TSC organizers.

4.4 Using Query Words from Questions

We used the questions given by the NTCIR-4 TSC organizers in two ways. The first was to order clusters so that they corresponded to queries extracted from questions. The other way was to weight sentences according to the queries. If queries were not given, we could substitute total word frequencies in all source documents for the queries. This result is shown in Table 9. Of course, our coverage decreases slightly by 0.02 to 0.03 points, but the coverage without queries (0.337) still ranked second, as seen in Table 2. In addition, redundancy defined by the difference between coverage and precision was drastically reduced. We also showed the responsiveness evaluation that is the average of long and short summary results. The exact match result was not so good because we did not use sentence type information for this result, but the edit distance result still shows better values.

5 Sentence-type Filtering

We experimented to see whether the sentence-type filtering approach was effective in improving responsiveness for short single-document summarization [17]. Sentence-type filtering reflects functional aspects of documents that are orthogonal to topical aspects using content words. We also tried the sentence-type filtering approach for multi-document summarization. [11] claimed that this approach was helpful in providing context to users of summarization in the medical domain. We show our annotation framework in this section.

5.1 Sentence type Annotation

We defined five sentence types as originally proposed in [8]: "Main Description", "Elaboration", "Background", "Prospectives", and "Opinion". Two assessors annotated these five types manually to 604 Nikkei newspaper articles published over two days in 1994. At inter-coder sessions, they gave each sentence the commonly recognized type. Of the 604 articles, we excluded 357 articles that contained only "Main Description", "Elaboration" and "Background" sentence types. The remaining 247 articles contained 4015 sentences.

Machine learning such as SVM could not be applied directly to identify sentence type because there were too few annotated "Opinion"- or "Prospective"-type sentences. We established type-oriented clues: for example, heading word counts in a sentence for "Main Description" type, opinion-oriented modal verbs, or background-related data suffix units. We defined about 100 clues.

5.2 Sentence-type Filtering

We tried to use sentence-type information to improve the responsiveness to questions given by organizers. Our system's strategy was based on a two-sentence extraction strategy as follows.

1. The most heavily weighted sentence in each cluster was extracted.
2. For the second or third weighted sentence in each cluster, the sentence-type information was checked.
   (a) The redundancy of sentence type for the most weighted sentence in the same cluster was checked first.
   (b) Then, if the sentence type was "Opinion" or "Prospective", we extracted it to produce summaries.
   (c) No more than two sentences were extracted from any cluster.

This strategy improved the responsiveness for some topics. The results are shown in Table 10, which only shows the improvement for exact matches.

6 Conclusions

In this NTCIR-4 TSC, we focused on multi-document summarization from two different aspects: topical and functional differentiation. We implemented topical aspects differentiation using document-clustering techniques. We considered paragraph as the basic cluster unit and their feature vectors were computed based on non-normalized term frequency, because they were rather small. Using this technique, our system attained high coverage and low precision for the extract evaluation but reduced redundancy successfully.

We implemented a summarization system interface called *SWIM* that could answer the user requirements, as shown in Figure 1. Users could specify their information requirements not only by queries from topical aspect, but also by requesting a "summary type." We call this the "orthogonal aspect." In this research, we implemented part of this aspect as sentence-type information to improve the responsiveness to questions.

Table 9. Coverage, Precision, and Responsiveness Change using Query Words and Total Frequency Words

Algorithm		Ward using Queries			Ward using Total Frequencies		
Distance		Euclidean					
Unit		Paragraph					
Feature		Term frequency vectors for nouns and unknown words					
Cluster Number for Long Summaries (× Number of Sentences Extracted)		× 1	× 1.5	× 2	× 1	× 1.5	× 2
Cluster Number for Short Summaries (× Number of Sentences Extracted)		× 1.5	× 2	× 2.5	× 1.5	× 2	× 2.5
Extract one sentence from each cluster	Coverage	**0.364**	0.357	0.353	**0.337**	0.321	0.333
	Precision	**0.518**	0.543	0.565	**0.450**	0.454	0.498
	Redundancy (=Precision-Coverage)	**0.154**	0.161	0.212	**0.113**	0.133	0.165
	Responsiveness (exact)	0.331	0.334	0.340	0.257	0.275	0.272
	Responsiveness (edit)	0.721	0.726	0.727	0.701	0.702	0.695

Table 10. Responsiveness Improvement with Sentence-type Filtering

ID	Topic	L/S	Sentence Type Not Used		Sentence-type Filtering		
			Responsiveness				Type
			exact	edit	exact	edit	
0310	Fossil in Ethiopia	L	0.200	0.780	0.300	0.798	Prospective
0410	Nakata movement (Soccer)	S	0.273	0.861	0.364	0.854	Prospective
0450	Company subsidiary move	L	0.214	0.758	0.286	0.770	Prospective
0510	Neutron	S	0.444	0.847	0.556	0.875	Prospective
0560	Mistake in entrance examination	L	0.545	0.942	0.636	0.942	Prospective
0570	Space Shuttle	S	0.308	0.835	0.385	0.843	Prospective
0630	Ancient tomb	L	0.364	0.849	0.455	0.866	Opinion

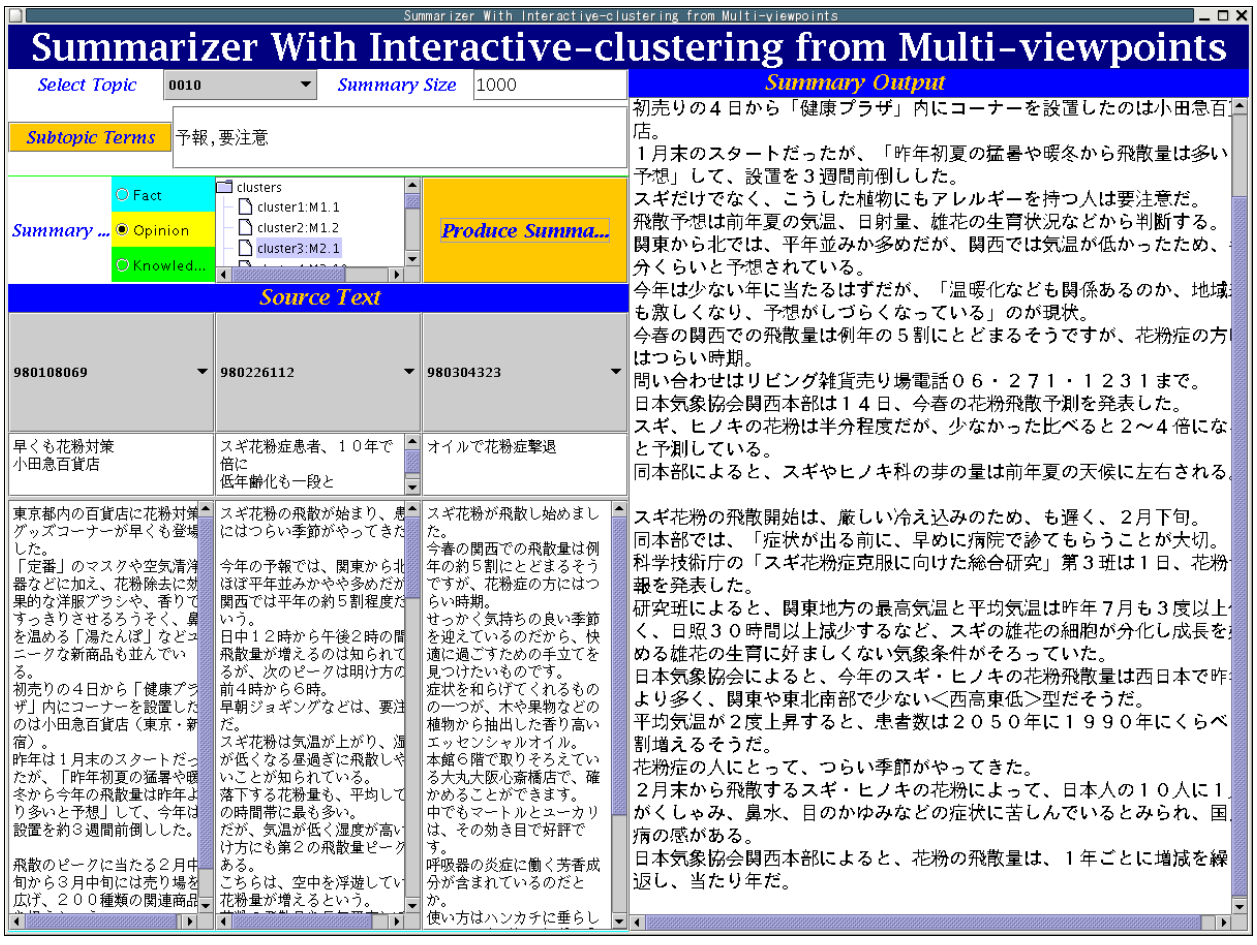

Figure 1. Multi-Document Summarization System *SWIM* for Two Aspects of Information Requirements

References

[1] R. Angheluta, R. De Busser, and M.-F. Moens. The use of topic segmentation for automatic summarization. In *Workshop on Text Summarization (DUC 2002) at the Association for Computational Linguistics 40th Ann. Meeting (ACL 2002)*, Philadelphia, Pennsylvania, July 2002.

[2] E. Boros, P. B. Kantor, and D. J. Neu. A clustering based approach to creating multi-document summaries. In *Workshop on Text Summarization (DUC 2001) at the ACM SIGIR Conf. 2001*, New Orleans, Louisiana, Sep 2001.

[3] A. El-Hamdouchi and P. Willett. Hierarchic document classification using Ward's clustering method. In *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 149–156, Palazzo dei Congressi, Pisa, Italy, 1986. ACM Press.

[4] W. B. Frakes and R. Baeza-Yates, editors. *Information Retrieval Data Structures & Algorithms*. Prentice Hall, 1992.

[5] T. Fukusima, M. Okumura, and H. Nanba. Text Summarization Challenge 2: Text summarization evaluation at NTCIR Workshop 3. In *(Oyama, Ishida, & Kando 2003)*, 2003.

[6] V. Hatzivassiloglou, J. L. Klavans, M. L. Holcombe, R. Barzilay, M. Y. Kan, and K. R. McKeown. Simfinder: A flexible clustering tool for summarization. In *Workshop on Automatic Summarization at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, pages 41–49, Pittsburgh, Pennsylvania, June 2001.

[7] T. Hirao, M. Okumura, T. Fukusima, and H. Nanba. Text Summarization Challenge 3: Text summarization evaluation at NTCIR Workshop 4. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question Answering, and Summarization*, 2004.

[8] N. Kando. Text Structure Analysis Based on Human Recognition: Cases of Japanese Newspaper Articles and English Newspaper Articles (in Japanese). *Research Bulletin of National Center for Science Information Systems*, 8:107–126, 1996.

[9] N. Kando. Overview of the Third NTCIR Workshop. In *(Oyama, Ishida, & Kando 2003)*, 2003.

[10] I. Mani. *Automatic Summarization*, volume 3 of *Natural Language Processing*. John Benjamins, Amsterdam, Philadelphia, first edition, 2001.

[11] L. McKnight and P. Srinivasan. Categorization of sentence types in medical abstracts. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*, pages 440–444, Ottawa, Canada, November 2003.

[12] M.-F. Moens. *Automatic Indexing and Abstracting of Document Texts*. Kluwer Academic Publishers, 2000.

[13] K. Oyama, E. Ishida, and N. Kando, editors. *Proceedings of the Third NTCIR Workshop on research in information Retrieval, Automatic Text Summarization and Question Answering*. National Institute of Informatics, 2003.

[14] D. R. Radev, H. Jing, M. Stys, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938, 2004.

[15] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 1988.

[16] Y. Seki. Sentence extraction by TF/IDF and position weighting from newspaper articles. In *(Oyama, Ishida, & Kando 2003)*, 2003.

[17] Y. Seki, K. Eguchi, and N. Kando. Compact summarization for mobile phones. In *Mobile and Ubiquitous Information Access, Lecture Notes in Computer Science 2954*, pages 172–186. Springer-Verlag, 2004.

[18] G. C. Stein, T. Strzalkowski, and G. B. Wise. Summarizing multiple documents using text extraction and interactive clustering. In *Pacific Association for Computational Linguistics (PACLING-1999)*, 1999.

[19] G. C. Stein, T. Strzalkowski, G. B. Wise, and A. Bagga. Evaluating summaries for multiple documents in an interactive environment. In *Proceedings of the Second International Conference on Language Resources & Evaluation (LREC2000)*, pages 1651–1657, Athens, Greece, 2000.

[20] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

[21] M. Zhang, C. Lin, Y. Liu, L. Zhao, and L. Ma. THUIR at TREC 2003: Novelty, robust, web and HARD. In *Proceedings of the Twelfth Text Retrieval Conference (TREC) 2003*, Gaithersburg, MD, November 2003. NIST Special Publication 500-252.