A Query-Free Retrieving Method Based on Content Elements' Order for Multimedia News Archives

Daisuke Kitayama and Kazutoshi Sumiya

School of Human Science and Environment, University of Hyogo
1–1–12 Shinzaike-honcho, Himeji, Hyogo 670-0092, Japan
ne07p001@stshse.u-hyogo.ac.jp, sumiya@shse.u-hyogo.ac.jp

Abstract. Video and text-news content have recently been broadcast on TV, newspapers, and the Internet. Although video content on out-of-date news is of little value for viewing, it can be considered to have value by comparing it to related content. Repeated news should especially be compared, e.g., the Olympic games and international expositions. We propose a method of retrieving comparison content based on the order of news elements. It is composed of two parts. The first is an analysis of news content that someone is browsing. The second is the automatic generation of queries for retrieving content on comparison news.

1 Introduction

Information is generally distributed by the news not only on TV and by newspapers but also by the Internet. However, the news is only reported on these sites short term (about a week or a month). The currency of the news from these sites is generally assumed to be important. Articles that compare past Olympic events with present ones, on the other hand, are composed of feature articles. Old news that is not browsed is not considered to be of any value. However, we considered the value by finding the relation between past and present news.

The method we propose has two processes. First, objects or things that have been reported and news behaviors (actions) are extracted based on the order of content elements, which differs depending on the media. Second, news articles are retrieved that can effectively be compared with the news article that the user is browsing. A user can automatically obtain content to understand the news with our method simply by browsing news articles and selecting the comparison query. Figure 1 outlines the concept underlying the method we propose.

2 Related Work

Shin et al. [1] proposed generating query from natural language questions. Their proposed systems automatically analyzed a user's question using the 5-W 1-H keywords. Our proposed method needs neither complex grammatical analysis

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 216–219, 2007.

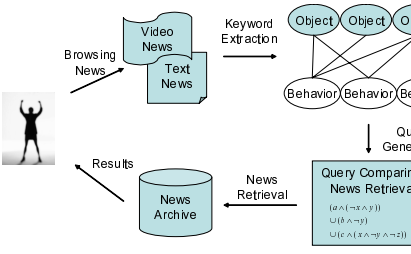

Fig. 1. Concept underlying query-free comparative news retrieval

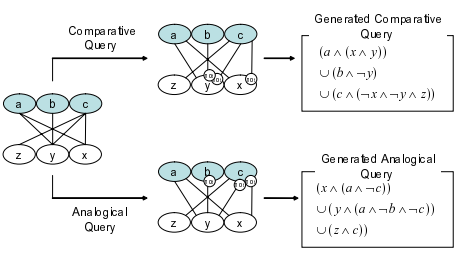

Fig. 2. Generation of comparison queries

nor dictionary building because it does not depend on specific keywords. Instead the composition keywords of the news are simply extracted.

Ohshima et al. [2] proposed methods of extracting the sibling page. Yumoto et al. [3] proposed methods of extracting relational page sets. They proposed a method of detecting the relations between content using a vector space model. The relations with our method are detected based on keywords without using a vector space model.

3 Keyword Extraction Using Order of Content Elements

We defined the order of content elements as elemental units in the order of news content. They have different features based on media[4][5]. We consider that the objects of subjects described in news are often expressed as nouns, and behaviors are expressed as sets of verbs. One news item can be expressed by using the noun for the object and the verb for the behavior.

The method we propose extracts objects from subjects described in the news using elemental units. We consider that the object of a subject in video news accurately describes the object spotted at the scene. In this way, the degree of importance of the object keyword can be calculated from the word density in the transcription of video news. The degree of importance of the object keyword, a, in video news can be calculated as

$$obj\_val = \frac{i}{dist(a_1, a_i)} \tag{1}$$

where a_i is the i^{th} noun a in the news. Function $dist$ calculates the distance between sentences. The distance between sentences is represented by a number and means how many sentences there are between two keywords. The distance between sentences is 1 when they appear in the same sentence. We consider that the positions where the objects of subjects described in text news are dispersed. The degree of importance of the object keyword, a, in text news is calculated as

$$obj\_val = min\left(\frac{\sum_{i=1}^{n} dist(s_1, a_i)}{n}, ..., \frac{\sum_{i=1}^{n} dist(s_j, a_i)}{n}, ..., \frac{\sum_{i=1}^{n} dist(s_m, a_i)}{n}\right) \tag{2}$$

where s_j is the j^{th} sentence in text news. The minimum value of the element is extracted using function *min* because the expectation is unknown.

The method we propose extracts news behaviors using the order in which content is presented. We considered the conclusion to be described at the end of video news, and the verb that shows action in the conclusion expressed the news behavior. The degree of importance of the behavior keyword is calculated by the position it appears in the transcription of video news. The degree of importance of the behavior keyword in video news is calculated as

$$beh\_val = \sum_{i=1}^{S}(\frac{i}{S} \times count(V_i)) \tag{3}$$

where i is the i^{th} sentence in all S sentences, and V_i is a verb set that appears in the i^{th} sentence. Function *count* calculates the number of verbs to be calculated in V_i. We considered that a news-behavior keyword in text news would appear at the beginning where details on the conclusion are described. The degree of importance of the behavior keyword in text news is calculated as

$$beh\_val = \sum_{i=1}^{S}(\frac{S-i+1}{S} \times count(V_i)). \tag{4}$$

4 Queries Generation for Retrieving Comparison Articles

Comparison articles are those which can be compared by focusing attention on the browsing news. We defined the news where the focus of attention was an object as a comparative article, and that where the focus was a behavior as an analogical article. The query is generated based on a graph where the content element is described. The content element graph is a bipartite composed of the object keyword and the behavior keyword. The link shows the relation between the object and the behavior. We consider that the relation between the object keyword and the behavior keyword in video news is determined by a range where the word density of an object keyword is high. However, the relation between the object keyword and behavior keyword in text news was determined using the same paragraph.

Comparative queries are automatically generated to retrieve comparative articles related to what the user is currently browsing. The user can confirm whether the situation has been consistent over time. The upper part of Figure 2 shows how a comparative query is generated.

Analogical queries are automatically generated to retrieve analogical articles related to what the user is now browsing. The user can understand behavior in detail. The lower part of Figure 2 shows the generation of an analogical query.

5 Evaluation

We did an experiment to evaluate our proposed method by assessing the retrieved results for generating queries using a news elements' graph. The data

Table 1. Experimental result of retrieving comparison article

Video News No.		Comparative			Analogical		
		Precision	Recall	F-measure	Precision	Recall	F-measure
1	Text	0.50	0.25	0.33	0.80	0.33	0.47
	Video	1.00	0.25	0.40	0.64	0.39	0.49
	All	0.67	0.25	0.37	0.68	0.37	0.48
2	Text	0.00	0.00	0.00	0.43	0.33	0.38
	Video	0.00	0.00	0.00	0.44	0.71	0.54
	All	0.00	0.00	0.00	0.43	0.57	0.49

Text News No.		Comparative			Analogical		
		Precision	Recall	F-measure	Precision	Recall	F-measure
3	Text	0.00	0.00	0.00	0.50	0.25	0.33
	Video	0.33	0.40	0.36	0.00	0.00	0.00
	All	0.22	0.18	0.20	0.27	0.14	0.19
4	Text	0.29	0.40	0.33	0.05	0.33	0.08
	Video	0.17	0.50	0.25	0.33	0.27	0.30
	All	0.20	0.43	0.27	0.13	0.29	0.18

sets for each generated query in the experiment were about 180 news items in the news archive[1]. Video news and text news were included in these data sets. Test subjects extracted correct-answer sets from data when browsing news that generated queries and then compared queries. There were three test subjects. A correct-answer set was a set of articles that two or more test subjects extracted. We evaluated the proposed method by means of precision, recall, and F-measure calculated using correct-answer sets.

The results are listed in Table 1. F-measure of query generated from video news is higher than F-measure of query generated from text news. We considered that different media are not equally processable in our proposed method. Therefore, we should improve the algorithm.

6 Concluding Remarks

A content element graph with the degree of importance based on the order of content elements was presented, and the generation of queries to compare articles that had been retrieved using this graph was proposed. We also evaluated the retrieved results using comparative queries generated with our proposed method. In future work, we plan to: compare methods of calculating keywords in experiments with conventional degrees of importance, and improve generation of queries based on the relations between individual keywords.

References

1. Shin, S.E., Seo, Y.H.: Query Generation Using Semantic Features. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 234–243. Springer, Heidelberg (2006)
2. Ohshima, H., Oyama, S., Tanaka, K.: Sibling Page Search by Page Examples. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 244–253. Springer, Heidelberg (2006)
3. Yumoto, T., Tanaka, K.: Page Sets as Web Search Answers. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 244–253. Springer, Heidelberg (2006)
4. Wikinews: Style guide, `http://en.wikinews.org/wiki/Wikinews:Style_guide`
5. Analizing News, `http://akasaka.cool.ne.jp/kakeru3/bs3.html`

[1] About 180 news items were assumed because 8 news items were used every month for 18 months and about 40 news items were selected by subjects as correct answers.