The Photo News Flusher: A Photo-News Clustering Browser

Tatsuya Iyota and Keiji Yanai

Department of Computer Science,
The University of Electro-Communications
1–5–1 Chofugaoka, Chofu-shi, Tokyo, 182–8585 Japan
{iyota-t,yanai}@mm.cs.uec.ac.jp

Abstract. We propose a novel news browsing system that can cluster photo news articles based on both textual features of articles and image features of news photos for a personal news database which is built by accumulating Web photo news articles. The system provides two types of clustering methods: normal clustering and thread-style clustering. It enables us to browse news articles over several weeks or months visually and find out useful news easily. In this paper, we describe an overview of our system, some examples of uses and user studies.

1 Introduction

Many commercial news sites exist on the Web, and they deliver a lot of new articles to us every day. Since data on the Web can be collected automatically by crawler programs, we can accumulate news articles on the Web automatically and build a personal news database with almost no cost. We can gather more than one thousand articles a month, so that it is very difficult to watch all of them and find out interesting articles out of such huge news database.

In this paper, we propose a novel news browsing system which can cluster photo news articles based on both textual features of articles and image features of news photos for a personal news database on PC. The system provides two types of clustering methods: normal clustering and thread-style clustering. For the normal clustering, we can adjust weights of textual features and image features, and for the thread-style clustering, we can control width of thread branches with a novel method. Ide et al. [1] extracted topic threads from a large-scale TV news video corpus and created a thread-based interface which enables users to browse news articles related to the same topic along time series. This topic threading is helpful to browse a large amount of news articles, so that we import and modify it for our system as one of the clustering methods. By these functions, the proposed system enables us to browse news articles over several weeks or months visually and find out useful news easily. In this paper, we describe an overview of our system, examples of uses, and user studies.

2 Proposed System

Regarding Web news search, it is general to search for news articles without photos using only textual information. On the other hand, our targets are news articles with photos, which have a special characteristic that it can be understood intuitively with just a look without reading. Our purposed system cluster news articles using both textual information from text news articles and visual information from photo articles. Since photos have the advantage of being visually recognizable, results of clustering are also easy to understand visually with just a look. We aim to achieve a system which enables us to find interesting news articles easily taking advantage of photos as visual clues.

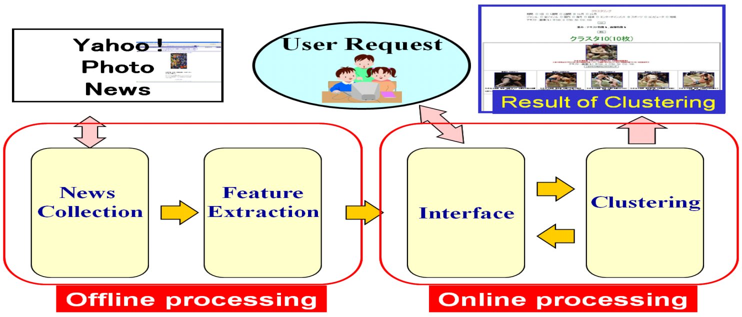


Fig. 1. Structure of the proposed system which consists of the four parts.

We collect news articles with photos from the Yahoo Photo News Japan. Being different from usual news articles, the main contents of articles of the Yahoo Photo News are photos. In addition to a photo, each article has a title and a short main text which explain the contents of the photo. Note that all articles are written in Japanese, since the Photo News site we use in our work is the Yahoo Photo News Japan. But our method did not rely on a specific language, so we can adjust our system for the other languages easily.

Our system consists of four parts: **News Collection Part**, **Feature Extraction Part**, **User Interface Part**, and **Clustering Part** (Fig.1). The News Collection Part and the Feature Extraction Part are carried out once a day as off-line preprocessing. The User Interface Part and the Clustering Part are carried out on-line.

News Collection In the News Collection Part, we gather titles, main texts, and photos of news articles by following the links from the top page of Yahoo Photo News. If the Web page of the photo news includes a link to a usual news article which has no photos, we collect the main text of the normal news article as a "linked text" of the photo news article.

Next, we classify news articles collected into seven categories (domestic, international, business, entertainment, sports, computer, and local). The seven categories are already classified by the Yahoo Photo News and can be distinguished from the URL of news articles. In addition, we prepare "all" to which all articles belongs.

Feature Extraction In the Extraction Part, we extract image features from the gathered photos and textual features from the titles, the main texts and the linked texts gathered in the Collection Part.

As image feature, we use color histograms computed in the Lu^*v^* color space as color image features. Each histogram quantizes the color space into 64 (4 for each axis) bins. In addition, we also use the bag-of-keypoints histogram [2] which consists of 300 bins as texture image features.

As textual feature, we use the vector space model for textual features. Each element of keyword vectors is weighted by the entropy-based TF-IDF. Similarity between articles is calculated as follows:

$$sim = weight * (sim_{img\_color} + sim_{img\_texture}) + (1 - weight) * sim_{text} \quad (1)$$

where sim_{img_color} is the similarity calculated based on color image features, $sim_{img\_texture}$ is the similarity calculated based on bag-of-keypoint-based texture image features, sim_{text} is the similarity calculated based on textual features, and $weight$ $(0 \leq weight \leq 1)$ is a weight of image features. $Weight$ is adjusted by a slider on the User Interface. Similarities based on both color and texture image features are calculated by the histogram intersection. Before calculating similarities, each feature vector is normalized so that the sum of its elements is 1. Similarity based on textual features is calculated by the cosine similarity.

User Interface (UI) A user selects a term and a category of news articles and set *weight* based on his/her preference on the Web-based UI. *Weight* is the ratio of the textual feature and the image feature used in similarity calculation for clustering. The Web-based user interface provides us with a slider to set a weight of textual features and image features.

Clustering In the Clustering Part, the system clusters news articles according to similarity of feature vectors extracted in the Feature Extraction Part, and shows the article which is closest to the mean of the cluster as the representative article of each cluster. We can see all the images of the cluster by selecting "display all images" on the UI.

The proposed system provides two kinds of clustering: k-means-based normal clustering and time-series-based thread clustering.

Thread clustering groups articles related to the same topic and show them in the time order. Here, "thread" means series of articles which are related to the same one topic in the time order. If several articles are not similar each other but all of them are similar to a certain article, "branching" will be made. "Thread" can includes branches, and is composed as tree-structure in general.

By thread-clustering, we extract a thread which starts from a certain topic a user selects from the result of normal k-means clustering, and show several articles arranged as a short comic strip. They enable us to find out the flow or relation of affairs or events more easily. In the UI, a user also select a constant α which decides what extent of topic drift in a thread is allowed. "Topic drift" means that the main topic of a thread is changing gradually along time progress. This "topic drift" control by a constant is one of novelties of our system, compared to Ide et al. [1].

To find child articles which are highly related to the parent and their time stamps are always newer than the parent's. "Highly related" is defined by the following condition:

$$sim(v_n, V_{n-1}) \leq T \tag{2}$$

, where

$$V_n = \begin{cases} v_n & (n = 0) \\ v_n + \alpha V_{n-1} & (n \geq 1) \end{cases} \tag{3}$$

$$\left(\begin{array}{ll} \alpha : \text{a constant which controls "topic drift"} & (0 \leq \alpha \leq 1) \\ T : \text{a threshold} & \\ v_n : \text{a feature vector of } n\text{-th article} & \\ V_n : \text{an } n\text{-th accumulative mean vector} & \end{array} \right)$$

3 Example Uses and a User Study

Use of Normal Clustering We assume a user have an interest on the Japan national team of World Cup of soccer. At first, the user selects "two months" as the term of news articles and "sports" as the category. Next, the user selects "text 100%" as the weight of textual and image features so that the clusters are associated with a sub-category of sports such as soccer and golf (Fig.2). The representative image in the third cluster is soccer, so we can estimate the third cluster contains many soccer articles.

In the next step, the user carries out re-clustering for the third cluster with text 50% and image 50% in order to divide soccer articles into some clusters. As a result, the user find out a "Japan Soccer W-cup" cluster which includes a lot of articles related to the Japan National Soccer Team (Fig.3).

Use of Thread Clustering We assume that the thread clustering is used for searching for related articles along time series after a user find out an interesting article by the normal clustering.

Fig.4 shows the thread on the food poisoning affair in the Japanese Domestic news which was one of biggest news in early 2007 in case that α is 0, while Fig.5 shows the thread on the same topic in case that α is 0.9. These results indicates the extent of branching is adjusted by varying the value of α.

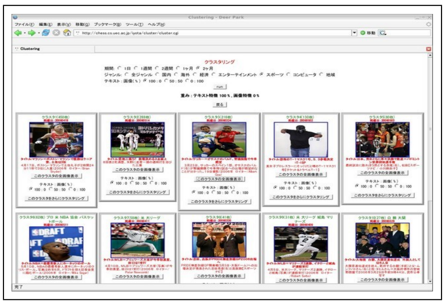
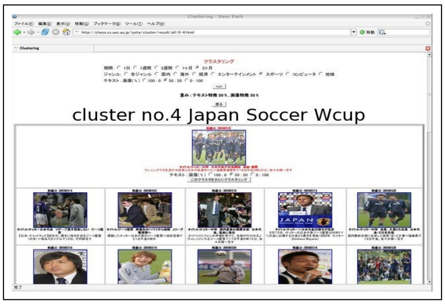
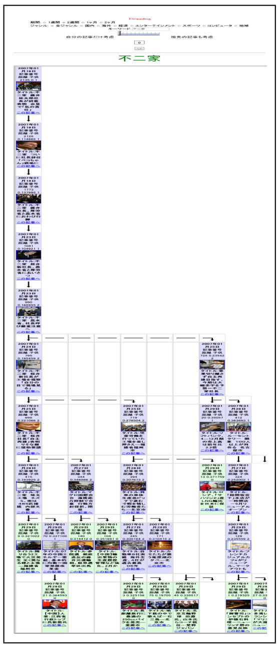


Fig. 2. Clustering result of sport news articles.



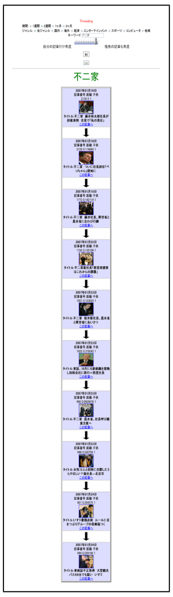
cluster no.4 Japan Soccer Wcup

Fig. 3. "Japan Soccer W-cup" cluster.



不二家

Fig. 4. Thread ($\alpha = 0$).



不二家

Fig. 5. Thread ($\alpha = 0.9$).

User Study We made the brief user study on the proposed system. We asked seven subjectives to compare the proposed system with a baseline system regarding how easily to search for interesting articles based on their own preferences. The interface of the baseline system is similar to the Yahoo Photo News site. The subjectives evaluated both the baseline and the proposed system with a score from 1 to 5. As a result, the proposed system and the baseline system obtained 3.86 and 2.43 on average, respectively. By applying Student's T-test, the difference on the average scores was proved to be significant.

4 Conclusions

We proposed a new photo news clustering browser which provided normal and thread clustering. Its effectiveness was proved by the user study.

References

1. Ide, I., Mo, H., Katayama, N.: Threading news video topics. In: Proc. of ACM SIGMM MIR (2003) 239–246.
2. Csurka, G., Bray, C., Dance, C.R., Fan, L.: Visual Categorization with Bags of Keypoints. In: Proc. of ECCV WS on Stat. Learn. in CV, (2004) 1–22.