Color Recommendation for Text Based on Colors Associated with Words

Saki Iiba*, Tetsuaki Nakamura*, and Maki Sakamoto*

Abstract In this paper, we propose a new method to select colors representing the meaning of text contents based on the cognitive relation between words and colors. Our method is designed on the previous study revealing the existence of crucial words to estimate the colors associated with the meaning of text contents. Using the associative probability of each color with a given word and the strength of color association of the word, we estimate the probability of colors associated with a given text. The goal of this study is to propose a system to recommend the cognitively plausible colors for the meaning of the input text. To build a versatile and efficient database used by our system, two psychological experiments were conducted by using news site articles. In experiment 1, we collected 498 words which were chosen by the participants as having the strong association with color. Subsequently, we investigated which color was associated with each word in experiment 2. In addition to those data, we employed the estimated values of the strength of color association and the colors associated with the words included in a very large corpus of newspapers (approximately 130,000 words) based on the similarity between the words obtained by Latent Semantic Analysis (LSA). Therefore our method allows us to select colors for a large variety of words or sentences. Finally, we verified that our system cognitively succeeded in proposing the colors associated with the meaning of the input text, comparing the correct colors answered by participants with the estimated colors by our method. Our system is expected to be of use in various types of situations such as the data visualization, the information retrieval, the art or web pages design, and so on.

Key Words : Text Processing, Associative Color, Words, Color Recommendation.

1. Introduction

In the psychological field, processing the cognitive tasks is activated by external representations as physical symbols [19]. Color is a vital representation in the successful delivery of information, and appropriate colors lead to a profound understanding in text processing [8], [9]. Color information in the document is unconsciously considered essential for helping a good understanding of the text contents, for instance, to grasp the outline of the document quickly. In addition, it has been showed that the colors in the document are effective to memorize and recognize the text contents ([3], [10], [16]).

In this paper, we pursue the possibility of utilizing a visual indication accurately representing the meaning of textual information. From this viewpoint, we focus on proposing colors, which have the cognitive association with the text contents, to convey and strengthen the message from the textual information.

A previous study [14] clarified a mechanism of color association with music lyrics, and revealed the existence of specific words which had an effect on the colors associated with the meaning of text

* Graduate School of Informatics and Engineering, The University of Electro-Communications, JAPAN, sakamoto@hc.uec.ac.jp

contents. Taking account of the result from this study, we designed a method for estimating colors associated with the text contents based on the associative relation between words and colors. The goal of our study is to propose an interface which recommends colors plausible for the meaning of the input text.

2. Related Works

In the field of psycholinguistics, many studies have conducted various experiments to investigate the semantic relation between colors and associated verbal descriptions such as color terms [12] and emotional words [13], [15]. Cymbolism[1] is a web site that offers guidance about color symbolism. There exist several works on the automatic methods to determine colors associated with words depending on the knowledge based or corpus based semantic similarity between words and colors [11], [17].

To evoke the desired emotional responses, the practical design applications are developed. They estimate colors corresponding to emotional words in the input text and decorate the background or fonts in the text [7], [18]. The methods for visualizing textual information have been proposed based on the word color association ([4], [5], [6]). The colors representing the text contents are selected by the colors associated with the crucial words extracted from the text.

The most of previous studies, however, assign a few colors to a word. Since a small number of colors are applied to each word, the number of colors associated with the text is inevitably small. It is not sufficient to represent the meaning of the text contents, and the selected colors are hard to grasp details of the text contents. Therefore we employed 35 color samples, and defined the associative probability of each color corresponding to a word.

Although a large number of researches have been carried out into colors associated with words, little is known about strength of color association with a word. For example, *swan* is in general strongly associated with color white, on the other hand, it is difficult to associate colors with *justice*. We investigated not only colors associated with words, but also how strongly the colors are associated with each word. Therefore our method guarantees a strong association between words and colors, and this leads a close relation between text contents and colors.

In the previous studies, limited types of words such as color terms or emotional words have been employed. In contrast, we utilize the words that participants judged as strongly associated with colors, in other words, we set no restriction on the types of words. Furthermore, we make use of Latent Semantic Analysis (LSA) [2] which is a technique for obtaining the semantic similarity between words in corpus. Colors associated with words and strengths of color association with words are obtained by the similarity. Our method makes it possible to deal with a large variety of words or sentences.

Considering the colors associated with the word and the associative strength of each color, we estimate the probabilities of colors associated with text contents. The goal of our study is to propose a system to recommend the cognitively plausible colors for the meaning of the input text.

3. Text Color Estimation Method

Our method selects colors associated with the meaning of text contents based on the words which have the strong association with colors (hereafter,

[1] http://www.cymbolism.com/about

we call those words as "primitive words"). Each primitive word has the associated colors (with the associative probability of each color) and the strengths of color association, and we estimate the probabilities of colors corresponding to a text depending on the primitive words in the text.

3.1 Color Vector

In our study, 35 color samples in RGB system were employed as shown in Figure 1.

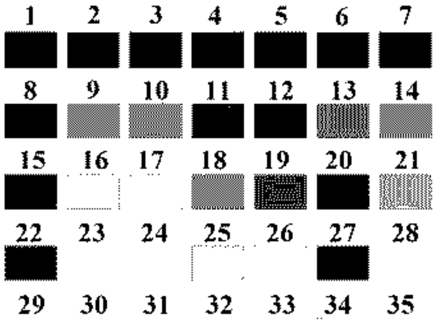


Figure 1. Color samples in our method.

Before describing our method, we clarify the concept of *color vector*. We define a color vector as the probability of each color associated with a word, namely a probability p_i of color c_i (i=1,⋯,35) associated with a word w, a color vector $v(w)$ of the word w is represented by the following equation:

$$v(w) = (p_1,\ldots, p_{35}) \tag{1}$$

For example, when w has a high value p_{15}, w is likely to be associated with color #15 as shown in Figure 1. Therefore we estimate the text color vector (i.e., the probabilities of colors associated with the text) by adding up all color vectors of primitive words extracted from the text.

3.2 Basic Method

In order to estimate the color vector for a text, we normalize the center of gravity vector, which is the sum total of the color vector of each primitive word. Each primitive word has the strength of color association, and we multiply the color vector of each primitive word by the strength of color association with the primitive word, that is, when a word has the strong color association, the associative probabilities of colors with the word are high values. Hence, the estimated colors for the text by our method are affected the colors of the primitive words which have the strong color association. A color vector $v(t)$ for a given text t is expressed by the following equation:

$$v(t) = \frac{\sum_{w_i \in A(t)} f(w_i) \cdot I(w_i) \cdot v(w_i)}{\left|A(t)\right|} \tag{2}$$

$A(t)$ represents the set of primitive words extracted from the text t, and w_i is a primitive word in the text t, therefore $w_i \in A(t)$. $f(w_i)$ and $I(w_i)$ indicate the frequency of appearance of the word w_i and the strength of color association with the word w_i. $v(w_i)$ shows the color vector of the word w_i.

4. System Design Procedure

We propose a system to recommend the colors plausible for representing the input text based on the color vectors associated with the primitive words. To estimate the colors for the input text, our method need to obtain three types of values related to the primitive words in the text: the word frequency, the strength of color association and the color vectors for each primitive word.

4.1 Overview of System

Figure 2 shows the overview of our system. The *user interface* takes the input text, and the *text analysis module* calculates the word frequency in the text using a morphological analysis tool[2]. Referring to the strength of color association and the color vectors for each primitive word in the *word database*, our system estimates the color vectors for the text in the *text color estimation module* applying our method (as described in Section 3.2). Finally, the *user interface* displays the colors based on the estimated color vectors.

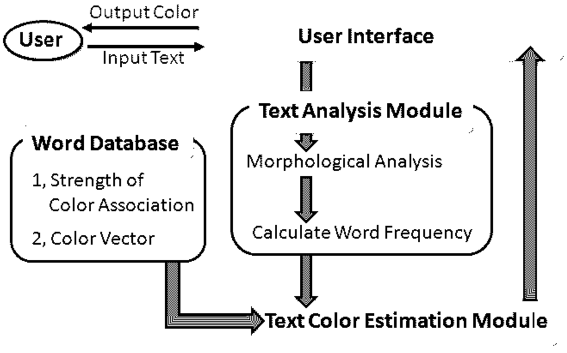


Figure 2. Overview of our system.

4.2 Database

Our system consists of two types of database: the strength of color association and the color vectors for primitive words. The data [14] extracted from music lyrics serve to examine our method. Therefore we applied the data [14] as basic data to our database. We conducted psychological experiments using news site articles to add more primitive words to our database.

4.2.1 Primitive Words

To find new primitive words from the news site articles, we conducted two psychological experiments in the same way as the previous study [14].

A) Experiment 1: Select Primitive Words

In this experiment, the participants read the text sample of news articles and answered colors associated with the text sample. Then we asked the participants to answer the words which strongly influenced the color association with the text sample. We considered the words answered by many participants as primitive words.

We decided to use news articles containing the words frequently appearing in general text documents, and selected 120 articles containing the words with high familiarity [1] within the articles published in 2009 by a news site organized by Mainichi Newspapers of Japan[3]. 80 Japanese undergraduates were divided into 4 groups, and we showed 30 articles per person. As a result, the data of 20 participants per text sample were obtained.

We asked the participants to answer the words influencing color association. The strength of color association of a given word indicates the rate of the number of participants who actually answered the word strongly influenced color association. Hence, the rate $R(w)$ of participants, who actually answered the word w has a strong color association, is represented by the following equations:

$$R(w) = \frac{\sum_{t_j \in T(w)} R(t_j, w)}{|T(w)|} \tag{3}$$

$$R(t_j, w) = \frac{n(t_j, w)}{n(t_j)} \tag{4}$$

t_j is a text sample containing the word w. $T(w)$ is the set of text samples containing the word w, therefore $t_j \in T(w)$. $n(t_j)$ indicates the number of participants who were shown the text sample t_j,

[2] Mecab0.98, http://mecab.sourceforge.net/

[3] Mainichi.jp, http://mainichi.jp/

and $n(t_j,w)$ is the number of participants who actually answered the word w has a strong color association. Therefore $R(t_j,w)$ shows the rate of participants who actually answered the word w has a strong color association in the text sample t_j.

We selected 498 words with the rate R of 0.25 or over (i.e., judged by 5 out of 20 participants that the word has a strong color association). Those rates R are regarded as the strength of color association I in equation (2). Consequently we selected 498 new primitive words and obtained the strength of color association of each primitive word.

B) Experiment 2: Obtain Color Vector

In this experiment, we investigated colors associated with the primitive words selected in the experiment 1. We asked the participants to answer colors associated with primitive words, and we obtained color vectors of primitive words.

The color samples were those given by Figure 1. We asked the participants to answer three colors, although the participants were allowed to answer the following three alternatives: answering the unique color three times [c_1, c_1, c_1], answering the unique color two times and another color [c_1, c_1, c_2], answering the three different colors [c_1, c_2, c_3]. 20 Japanese undergraduates were shown 498 primitive words. As a result, the answers given by 20 participants per primitive word were obtained.

We calculated a color vector of each primitive word. As described in Section 3.1, the color vector consists of the probability of each color associated with a word, namely the rate of the number of times which each color were answered by participants. Therefore a color vector $v(w)$ of a primitive word w is given by the following equation:

$$v(w) = \left(\frac{m(w, c_1)}{3x(w)}, \ldots, \frac{m(w, c_{35})}{3x(w)} \right) \quad (5)$$

$x(w)$ represents the number of participants shown the word w, and $m(w,c_i)$ shows the count of color c_i.

The data on the color vectors of 498 primitive words were registered in our database.

4.2.2 Unknown Word

The previous study [14] pointed out that it was impossible to investigate all the words strongly associated with colors by a limited extent of single experiment, and proposed the way to estimate the strength of color association and the color vector of the words which could not be investigated by their psychological experiment (hereafter, we call those words as “unknown words”).

We applied Latent Semantic Analysis (LSA) [2] which is a technique for obtaining the semantic similarity between words by means of a very large corpus. In our study, we employed the corpus of a Japanese major newspaper “Mainichi Shin bun” in 2005 (643,807 documents, 129,462 words).

Unknown words are selected from the words having the similarity with the primitive word over a threshold. The strength of color association and the color vector for the unknown words are estimated by the similarity with the primitive word. To be concrete, the data on unknown words are obtained by multiplying the data on the primitive words, which have the similarity with the unknown word over the threshold, by the similarity. Therefore a strength of color association $I(u)$ of a given unknown word u and a color vector $v(u)$ of the word u are given by the following equations:

$$I(u) = \frac{\sum_{w_i \in A(u,\theta)} s(u, w_i) \cdot I(w_i)}{|A(u, \theta)|} \quad (6)$$

$$v(u) = \frac{\sum_{w_i \in A(u,\theta)} s(u, w_i) \cdot v(w_i)}{|A(u, \theta)|} \quad (7)$$

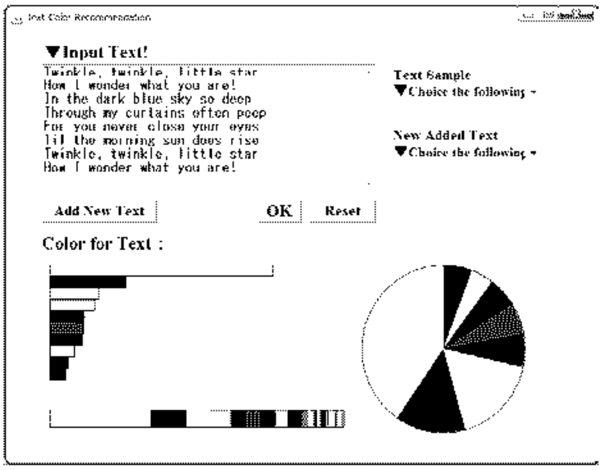
$$A(u, \theta) = \{ w_i \mid w_i \in A \wedge s(u, w_i) \geq \theta \} \quad (8)$$

A is the set of all primitive words. $I(w_i)$ and $v(w_i)$ show a strength of color association and a color vector of a primitive word $w_i \in A$. $s(u,w_i)$ indicates a similarity between the unknown word u and the primitive word w_i, and $P(u,\Theta)$ is the set of primitive words which have the similarity $s(u,w_i)$ over a threshold Θ .

As a result, we utilize the estimated data on the unknown words in approximately 130,000 words. Those data, in the same way as the data on primitive word, could be applied to our method as described in Section 3.2.

5. Prototype System

Figure 3 shows the interface of our prototype system. The system is implemented as a Java[4] program. User inputs text sentences in the box upper left in Figure 3, and by pressing the bottom "OK", the colors associated with the meaning of the text are displayed.



```
Twinkle, twinkle, little star
How I wonder what you are!
In the dark blue sky so deep
Through my curtains often peep
For you never close your eyes
Til the morning sun does rise
Twinkle, twinkle, little star
How I wonder what you are!
```

Figure 3. User interface of our prototype system.

6. Evaluation

To verify the efficiency of our method, we compared the colors proposed by our system (hereafter, we call those as "estimated color vector") with the colors answered by participants (hereafter, we call those as "correct color vector").

We started by conducting an experiment to obtain the correct color vector. In this experiment, the participants read the text samples of news articles, and we asked them to answer colors associated with the texts. 20 text samples were prepared, and the color samples were those in Figure 1. We offered the participants to answer colors in the same way in the experiment 2 (Section 4.2.1). 10 Japanese undergraduates were shown 20 text samples. As a result, the answers given by 10 participants per text sample were obtained. We calculated a color vector of each text sample applying the equation (5).

To compare the estimated color vector with the correct color vector, we analyzed by using a Pearson correlation coefficient between them. Table 1 shows the number of text samples having a combination which meets the significance level of 5%. From this table, in 65% (13 text samples, the mean correlation coefficient: 0.58) of all the text samples, the colors estimated by our system were correlated with the colors answered by the participants. This result indicates that our method

<Table 1> Number of text samples with the correlation coefficient meets the significance level of 5 %.

Values	Interpretation	Number of samples (%)
0.0 to 0.2	Very low	0 (0.00)
0.2 to 0.4	Low	4 (20.0)
0.4 to 0.7	Moderate	6 (30.0)
0.7 to 1.0	High	3 (15.0)
Total		13 (65.0)

The number of text samples: N=20

[4] http://www.java.com/en/

succeeded in selecting the cognitively plausible colors to represent the meaning of text contents.

7. Results and Discussion

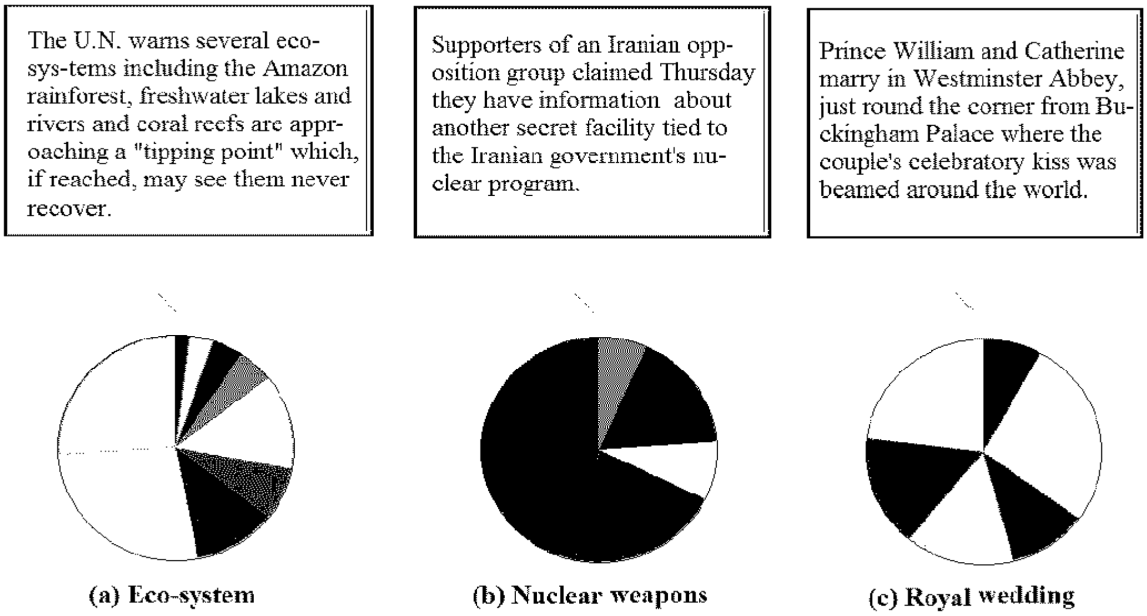
Figure 4 shows three examples of colors estimated by our system based on the color vectors of actual news articles in a news site[5].

We verified that our method succeeded in proposing the associative colors for the text in Section 6. In some text samples, however, unsuitable colors were proposed. To pursue a cause of this matter, we should examine our method with the various types of words or text documents and clarify whether the results from this study are Japanese specific or not. In our future work, we need to consider the more effective way to estimate color vectors applying the other data on the word color association obtained by [11], [17] or image search engines such as Google image search or Flicker. In addition, we are not concerned in our method with the topics (or latent semantics) from text contents. It will be necessary to incorporate such models into our method in order to extract commonsense knowledge and semantic concepts from text documents.

8. Conclusion

We proposed a method to select associative colors with the meaning of text contents based on the strong relation between words and colors. To estimate colors associated with text, we utilized the primitive words having strong association with colors. Each primitive word has the color vectors and the strength of color association. Our method estimates the color vector for a text, using the data related to primitive words.

The goal of this study is to provide a system to



The U.N. warns several eco-sys-tems including the Amazon rainforest, freshwater lakes and rivers and coral reefs are appr-oaching a "tipping point" which, if reached, may see them never recover.

Supporters of an Iranian opp-osition group claimed Thursday they have information about another secret facility tied to the Iranian government's nu-clear program.

Prince William and Catherine marry in Westminster Abbey, just round the corner from Bu-ckingham Palace where the couple's celebratory kiss was beamed around the world.

(a) Eco-system

(b) Nuclear weapons

(c) Royal wedding

Figure 4. Examples of colors plausible for texts proposed by our method

[5] CNN.com, http://www.cnn.com/

recommend the cognitively plausible colors for the text contents. We selected 498 primitive words and obtained the strength of color association and the color vector for each primitive word by conducting the psychological experiments. In addition, we employed the estimated data related to unknown words were derived from the similarities between words using LSA. Therefore our system can estimate a large variety of words or sentences.

Moreover we verified our method succeeded in proposing associative colors for the meaning of the texts. In our future work, we need to consider carefully whether our system is adoptable to various types of text documents, situations or languages.

References

[1] N. Amano, and T. Kondo, "NTT database series Nihongo Goitokusei [Lexical properties of Japanese]", *Sanseido*, Vol. 1 6, Tokyo Japan, 1999.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshma, "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391 407, Sept. 1990.

[3] R. H. Hall, and P. Hanna, "The impact of web page text background colour combinations on readability, retention, aesthetics and behavioural intention", *Behaviour & Information Technology*, Vol. 23, No. 3, pp. 183 195, May 2004.

[4] C. Havasi, R. Speer, and J. Holmgren, "Automated color selection using semantic knowledge". *In AAAI Fall Symposium Common Sense Knowledge*, Arlington USA, 2010.

[5] Y. Kiyoki, and X. Chen, "A semantic associative computation method for automatic decorative multimedia creation with 'kansei' information", *Proc. of the 6th Asia Pacific Conference on Conceptual Modeling*, Vol. 96, pp. 7 15, Wellington New Zealand, Jan. 2009.

[6] H. Liu, T. Selker, and H. Lieberman, "Visualizing the affective structure of a text document", *Proc. of the Conference on Human Factors in Computing Systems Computer Human Interaction*, Ft. Lauderdale USA, April 2003.

[7] C. Ma, H. Prendinger, and M. Ishizuka, "A chat system based on emotion estimation from text and embodied conversational messengers", *Proc. of the 4th International Conference on Entertainment Computing*, pp. 535 538, Kobe Japan, Sept. 2005.

[8] A. Marcus, "Color and communication: help is on the way", *ACM SIGDOC Asterisk Journal of Computer Documentation*, Vol. 15, No. 3, pp. 15 19, Nov. 1991.

[9] B. J. Meier, "ACE: A color expert system for user interface design", *Proc. of the ACM SIGGRAPH Symposium on User Interface Software*, pp. 117 128, Oct. 1988.

[10] C. B. Mills, and L. J. Weldon, "Reading text from computer screens", *ACM Computing Surveys*, Vol. 19, No. 4, pp. 329 357, Dec. 1987.

[11] S. Mohammad, "Colourful language: Measuring word colour associations", *Proc. of the ACL 2011 Workshop on Cognitive Modeling and Computational Linguistics*, Portland USA, June 2011.

[12] A. Mojsilovic, "A computational model for color naming and describing color composition in images", *IEEE Transactions on Image Processing*, Vol. 14, No. 5, pp. 690 699, May 2005.

[13] T. Nakamura, O. P. Sakolnakorn, A. Hansuebsai, P. Pungrassamee, and T. Sato, "Emotion induced from colour and its language expression", *Proc.*

of the AIC 2004 Color and Paints, pp. 328 331, Porto Alegre Brazil, Nov. 2004.

[14] T. Nakamura, K. Kawanishi, and M. Sakamoto, “A possibility of music recommendation based on lyrics and color” [in Japanese], *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, Vol. 94 A, No. 2, pp. 85 94, Feb. 2011.

[15] L. C. Ou, M. R. Luo, A. Woodcock, and A. Wright, “A study of colour emotion and colour preference”, *Color Research & Application*, Vol. 29, No. 3, pp. 232 240, June 2004.

[16] F. V. Scharff, A. L. Hill, and A. J. Ahumada, “Discriminability measures for predicting readability of text on textured backgrounds”, *Optics Express*, Vol. 6, No. 4, pp. 81 91, Feb. 2000.

[17] C. Strapparava, and G. Ozbal, “The color of emotions in texts”, *Proc. of the 2nd Workshop on Cognitive Aspects of the Lexicon*, pp. 28 32, Beijing China, Aug. 2010.

[18] K. Yamazaki., N. Muranaka, M. Sasajima, and N. Udagawa, “Mail system with considering kansei: Kansei mail”. *Annual design review of Japanese Society for the Science of Design*, Vol. 9, No. 9, pp. 52 57, March 2004.

[19] J. Zhang, and D. A. Norman, “Representations in distributed cognitive tasks”, *Cognitive Science*, Vol. 18, No. 1, pp. 87 122, Jan. 1994.