Proposal of the Kawaii Search System Based on the First Sight of Impression

Kyoko Hashiguchi and Katsuhiko Ogawa

Faculty of Environment and Information Studies, Keio University, 5322 Endo Fujisawa-shi, Kanagawa-ken, 252-0882, Japan
{t07624kh,ogw}@sfc.keio.ac.jp

Abstract. We propose a blog search engine called "Kawaii Search" (where Kawaii means pretty) to search blogs based on the impression of their text on a printing surface, considering factors such as the format and layout of text and density of words. Particularly in Japan, blogs reveal the personality characteristics of users depending on how they place their text. For example, some writers leave more space between lines or use hieroglyphics and "Gal words[1]," which consist of slang or abbreviations. Further, words can be categorized using four types of characters: kanji, hiragana, katakana, and alphabet. Each results in a different impression that reveals a writer's personality. Given this approach, blog readers can not only read blog, but also interpret each writer's personality. By focusing on impression differences, we propose a new search algorithm specialized for Japanese blogs. To show that these differences can act as the base of our search algorithm, we conducted an experiment that successfully verified the algorithm applied to the following three blog patterns: "kawaii" (pretty or lovely), "majime" (seriousness or industrious), and "futsu" (normal). The results show that in terms of the accuracy of the algorithm, our study categorized "kawaii" well; however, "majime" and "futsu" did not show good results.

Keywords: Impression, Blog search engine, text formatting, Japanese blogosphere, information retrieval.

1 Introduction

Blog search systems are generally based on the statistical and structural information in the blog text, including the frequency or relationships of words [2]. These systems search for blog articles that suit user requirements based on the content and keyword-based search techniques, including page rank and TFIDF [3].

With increase in the variety of blog writers' styles and readers searching blogs for different reasons, a more sophisticated search system is required. For example, users not only want to read an article that matches the content they are searching for, but also want to find a blog that meets their aesthetic requirements. Particularly in Japanese blogs, pictographs and emoticons are frequently used to express a blog writer's individuality.

Conventional search systems such as Google do not reveal the atmosphere or personality of its text; however, when people read blogs and diaries, they often look

G. Salvendy, M.J. Smith (Eds.): Human Interface, Part II, HCII 2011, LNCS 6772, pp. 21–30, 2011.

not only at the words, but also the design and layout of the text. The Kawaii Search system analyzes the qualitative information such as impression and layout of blogs quantitatively. We therefore propose this search system to find blogs which are visually preferred by users.

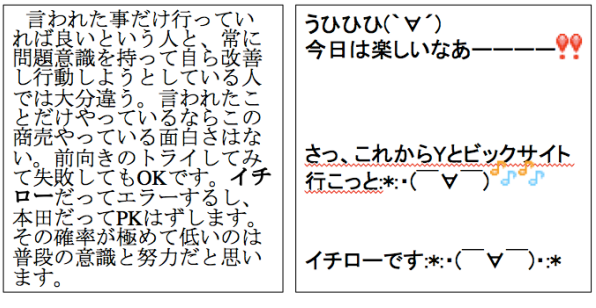
2 Kawaii Search

Kawaii Search is a system that searches blogs based on their appearance. In this section, we reveal typical differences found in blog appearances. We also describe the concept and vision of Kawaii Search.

2.1 Blogs in Japan

When we compare blogs in Japan to those in other countries, blogs in Japan have more character sets, including kanji, hiragana, katakana, and Roman. Further, Japanese blog writers use more spacing, symbols, pictograms, and emoticons than blog writers in other countries. These various styles give readers different impressions for each individual blog.

As an example, a blog discussing “Ichiro” and consisting of more kanji, less blank space, and no pictograms may give an impression of a serious blog, as shown in Figure 1(a). Conversely, even though the topic is the same, a blog that consists of less kanji, more blank space, and many of pictograms will have a pretty (or light) impression, as shown in Figure 1(b). Different writing patterns create different impressions.



(a) (b)

Fig. 1. Two blogs discussing “Ichiro”; (a) consists of more kanji, less blank space, and no pictograms; (b) consists of less kanji, more blank space, and many emoticons

In Japan, blogs written celebrities are popular among the public [4]. Each individual writer has a distinct writing pattern. Since writers know that many people read their articles, they may consider ways to increase readership through the blog’s textual layout. A writer may arrange the article by using effective spacing and emoticons so that it appears prettier, or use kanji for a smarter appearance. In other words, a blog’s

appearance indicates the impression of the article, and the style of writing reflects the writer's personality and character. Blog's readers not only understand the meaning of the blogs, but also read writer's personality and character. In this way, because a blog's appearance is important for reader and writer, we propose the search system based on the impression of their text on a printing surface, considering factors such as the format and layout of text and density of words.

2.2 Concept

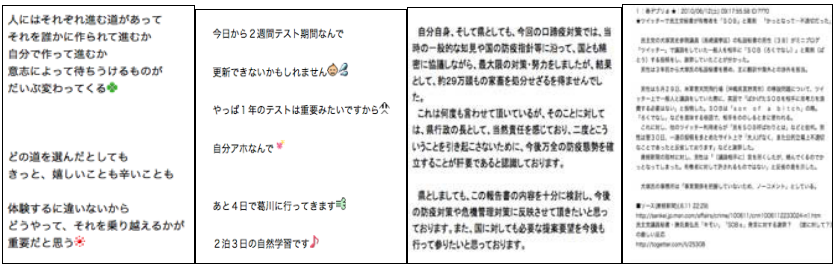
The Kawaii Search concept is described in this section. In addition to keyword-based blog search, Kawaii Search shows blogs similar to those of celebrities in terms of appearance. As illustrated in Figure 2, three celebrity blogs are shown at the top of the Kawaii Search site. These blogs are categorized by the following icons: Majime (serious), Kawaii (pretty), and Futsu (normal). Above these icons is a textbox in which users can enter their search keywords.

At the bottom of the Kawaii Search interface, search results are listed and the site shows the titles and thumbnails of each "hit." Instructions are as follows:

1. Enter the keyword and click one of the three icons shown (i.e., Majime, Kawaii, or Futsu).
2. The site will show blog articles similar to the one that the user selected. Figure 3 shows example results from a search request.

Fig. 2. Kawaii Search interface

In Figure 3, the first article is a kawaii celebrity blog, and the second article is a search result that is the top hits of kawaii. Since both writers use many pictograms and spaces, those articles have similar impression. Conversely, the third article is a majime celebrity blog, and the fourth article is the search result that is the top hits of majime. Both articles consist of more words and no pictograms or emoticons. As is evident from these examples, Kawaii Search can successfully match blogs that have a similar appearance to the base blog.



a. Kawaii celebrity blog b. Search result similar to kawaii celebrity blog c. Majime celebrity blog d. Search result similar to the majime celebrity blog.

Fig. 3. Example Kawaii Search results

3 The Kawaii Search Algorithm

3.1 Kawaii Value

It is difficult to quantify judgmental standards used when a person reads a blog article; however, in many cases, we may judge external characteristics of the blog using the overall impressions of sentences. This includes measures such as line row, character arrangement, condition of sentences, and so on. Kawaii Search focuses on these constituents, using the following six variables: Conspicuous Value, Words Value, Vertical Space Value, Hiragana Value, Emoticon Value, and Pictogram Value. An explanation of each of these follows.

Conspicuous Value. Conspicuous Value is based on how obvious Japanese characters are to the human eye. As illustrated in Figures 4(a) and 4(b), the conspicuousness of a word is based on the word itself and the words surrounding it. In Figure 4(a), when we compare 鸞 (phoenix) and 一 (one), 鸞 seems bolder, whereas in Figure 4(b), 鸞 in 鳥鸞 鳥 is not as prominent.

As described in Figure 4(a), when we compare "鸞" (phoenix) and "一" (one), "鸞" seems bolder. On the other hand, as shown in Figure 4(b), "鸞" in "鳥鸞 鳥" is not as bold as "鸞" in Figure 4(a).

We define conspicuousness as the Conspicuous Value based on the strokes of a character. The calculation method is as follows. As shown in Figure 5, the red boxes identify the groups of words. The blue numbers are the number of strokes in each word. For example, to calculate the Conspicuous Value of the word 晴れ (sunny) in the sentence shown, 今日は晴れです (it is sunny today), we first add 12 and 3, the number of strokes for 晴れ. Second, if we have multiple characters in the word, we calculate the average by dividing the resulting sum by the number of characters in the word (i.e., 15/2 = 7.5). Third, we add the number of strokes in words next to this word; in the example, the neighboring words are は and です (i.e., 4 + 4 + 3 = 11). Fourth, we divide this sum by the number of characters in those surrounding words (i.e., 11/3 = 3.67). Finally, we divide the average number of the word 晴れ by the

average number of the surrounding words (i.e., 7.5/3.67 = 2.04). The Conspicuous Value for 晴れ is therefore 2.04. By using these steps, all words are assigned Conspicuous Values.

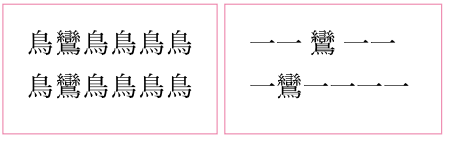


Fig. 4. Conspicuousness: (a) high conspicuousness; (b) low conspicuousness

Fig. 5. Example words and corresponding strokes for calculation of the Conspicuous Value of each word

Other Values. In this section, we describe the other proposed variables, i.e., the Words Value, Vertical Space Value, Emoticon Value, Pictogram Value, and Hiragana Value. Figure 6 illustrates these values using an example. Words Value is the number of words used in the given blog article. This value is used to measure the length of the article.

Fig. 6. Example blog excerpt with Words Value, Space Value, Emoticon Value, Pictogram Value, and Hiragana Value shown

Vertical Space Value, Emoticon Value, and Pictogram Value are the frequencies of appearance of those types of elements in the given blog article. Vertical Space Value corresponds to vertical space created by
 tags; Emoticon Value corresponds to emoticons; and Pictogram value refers to pictograms. Figure 7 shows examples of emoticons. These values quantify the characteristics of the article as follows:

$$\text{Emoticon Value} = \text{Emoticon} / \text{Words} \tag{1}$$

$$\text{Vertical Space Value} = \text{Space} / \text{Words} \tag{2}$$

$$\text{Pictogram Value} = \text{Pictogram} / \text{Words} \tag{3}$$

In each of the above equations, *words* refers to the number of words in the given blog.

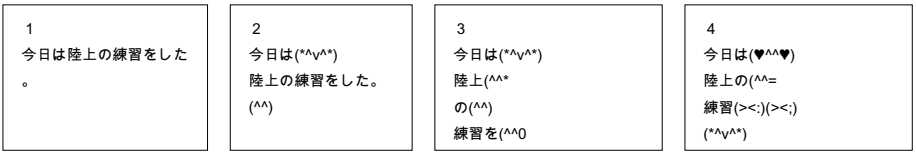


Fig. 7. Example blog excerpt ("I practiced track and field today") showing emoticons; in the first blog article, the Emoticon Value is zero; progressing left to right, the Emoticon Value. rise.

The Hiragana Value is the appearance frequency of hiragana in the given blog article, as illustrated in Figure 8. This value can be interpreted as the degree of softness, as expressed in the formula below (where *letters* do not include pictograms):

$$\text{Hiragana Value} = \text{hiragana} / \text{letters} \tag{4}$$

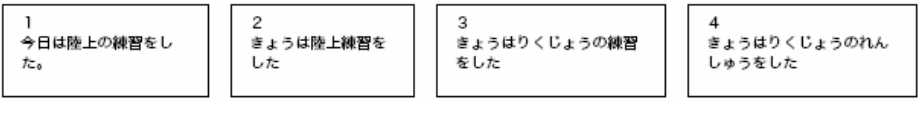


Fig. 8. Example blog excerpts illustrating hiragana; progressing left to right, the Hiragana Value rises

3.2 Blog Templates: Kawaii, Normal, Serious

We identified three blog templates as patterns for kawaii, futsu, and majime [5]. Each pattern's values are detailed below.

$$\textit{Majime Value} = 0.36 \times \textit{Conscious Value} + 0.33 \times \textit{Words Value} - 0.51 \times \textit{Vertical Space Value} \tag{5}$$

$$\textit{Kawaii Value} = 0.31 \times \textit{Vertical Space Value} + 0.26 \times \textit{Pictogram Value} + 0.22 \times \textit{Emoticon Value} + 0.16 \times \textit{Words Value} + 0.16 \times \textit{Conscious Value} - 0.25 \times \textit{Hiragana Value} \tag{6}$$

$$\textit{Futsu Value} = 0.28 \times \textit{Vertical Space Value} + 0.23 \times \textit{Emoticon Value} + 0.16 \times \textit{Hiragana Value} - 0.35 \times \textit{Words Value} - 0.06 \times \textit{Pictogram Value} \tag{7}$$

We set the Majime Value which can search the blog articles including a lot of numbers of characters, and a few spaces. Kawaii Value can search the blog articles including in a lot of spaces, pictograms and emoticons and words, and Futsu Value can pick out the blog articles including spaces, emoticons, and hiragana. We did scoring from the blog articles that the score is high by using these three values (Table 1).

Table 1. The characteristic of each Value

	Majime Value	Kawaii Value	Futsu Value
High score	*Conscious Value* *Words Value*	*Vertical Space Value* *Pictogram Value* *Emoticon Value* *Words Value Conscious Value*	*Vertical Space Value* *Emoticon Value* *Hiragana Value*
Low score	*Vertical Space Value*	*Hiragana Value*	*Pictogram Value* *Words Value*

3.3 System Structure

In this section we describe the Kawaii Search system, which is composed of three building blocks. Overall, the Kawaii Search system is implemented in PHP and MySQL. The first component is the crawler, which downloads the blog articles.

The second component is the indexer, which receives blog articles from the crawler and isolates the text. Next, the indexer analyzes the text using Mecab[6], which splits the text into its individual morphemes. In many cases, images, links, and advertisements are included in the given blog article. Our system does not accept any images, except for the pictographs, as determined by the algorithm. The indexer saves only the text, pictographs, emoticons, and
 (line breaks) in the underlying database, and it calculates the six values described above and stores them in the database.

The third component is the searcher. Users enter keywords and click on an icon categorized as Kawaii, Futsu, or Majime. The searcher obtains the six parameters for the blog article that the user clicked. At the same time, the searcher obtains blog articles in which the keywords match and retrieves the corresponding six parameters; this action occurs via the database. Next, the searcher calculates the difference between the acquired values and the values of the blog article that the user clicked. Scoring is done based on the number of differences (the fewer differences, the better). Finally, the searcher sorts in descending order and displays the results.

4 Evaluation

4.1 Experimental Method

By using our Majime, Kawaii, and Futsu indices, we conducted an evaluation experiment to assess whether the searched blog articles represent the personality of the writers.

The subjects in this evaluation were 12 college students in their 20s (6 women and 6 men); each participant reported reading blog articles before. In this time we set

searching words "Ukeru (*where Ukeru means many kinds of meaning such as interesting, receive, catch, fun and so on*)," "Aho(*where Aho means cluck)*," "Rikujyo (*Where Rikujo means track and field, land and so on*.)" which are able to search valorous blogs in terms of impression.

In the experiment, the top 10 (of 1200) blog articles that were searched using the Majime Value, Kawaii Value, and Futsu Value were used. Subjects were shown the blog articles which present blog articles that have been evaluated by many to be cute, normal, and serious. For each of the top 10 blog articles, subjects were asked to evaluate their similarity with the cute blog article, the normal blog article, and the serious blog article; the following 5-point scale was used: (5) very similar; (4) a little similar; (3) cannot say either; (2) not very similar; and (1) not similar at all.

4.2 Experimental Result and Discuss

Experimental Result. Figure 9 shows the mean values of the scores reported by the subjects. The upper side of the table lists the keywords used in the experiment. For example, searching the blog articles that contained the keyword "Aho" by using the Kawaii Value returned articles that many that people found cute, but only a few articles that were found serious or normal. In contrast, searching the articles by using the Futsu Value returned articles that were found to be normal and serious to almost the same extent.

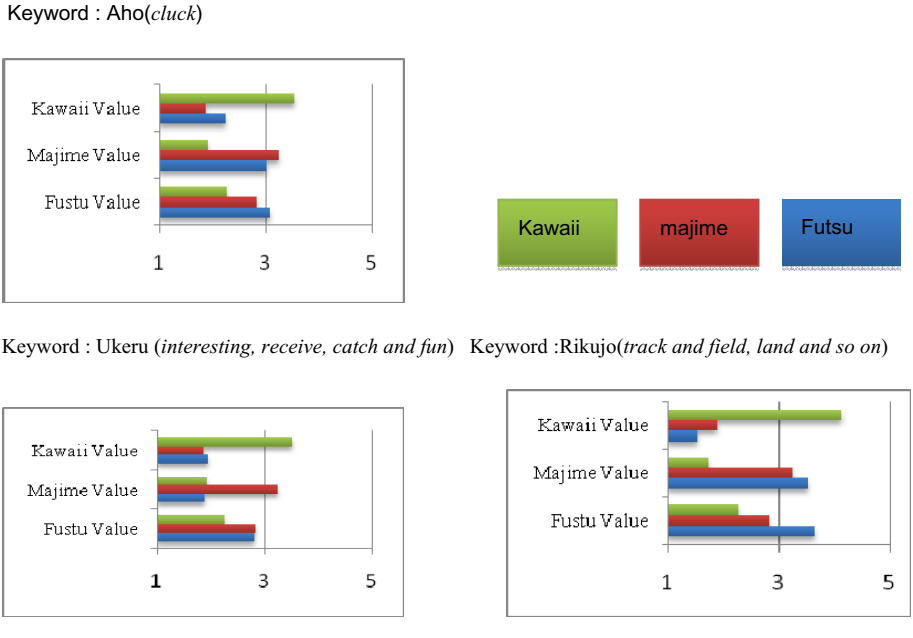


Fig. 9. Mean Kawaii Value, Majime Value, and Fujitsu Value reported by subjects

Discuss. As is evident in Figure 9, for any keyword, the Kawaii Value can be used to successfully select blogs that people find cute. The Majime Value and Futsu Value could not be used as they were similar. The Kawaii value is the indices that a blog article with many pictograms, spaces, emoticons is easy to be picked. Such blog articles are easy to understand the difference of the appearance in comparison with Majime and Futsu blog article. Therefore a blog article that user feel pretty is easy to be searched.

In the Majime value that picked out the articles including much number of words and little space between the lines. For search keyword “Ukeru,” results were successful. Because the word “Ukeru” has several possible meanings (i.e. receive, get, and fun), the crawler downloads many types of blogs. However, the other keywords could not be used as they were similar. It is thought that there was a problem for the articles that the crawl downloaded. For search keyword "aho" and “Rikujo”, space tends to become wide if the blog articles have much number of words. For example, in "Rikujo", there are many blog articles written by records of the time, and such a blog article has a lot of words and much space. In addition, with a “aho”, there are many articles transferred from the other sites such as twitter [7] or 2channel[8]. These blog articles have a lot of words and much space between the lines. Therefore, the blog articles with a little space between the lines are hard to be chosen.

Futsu Value could not be used as they were similar. There was a problem for the Futsu Value itself. The blog articles that are included many hiragana letters, and have much space are many kinds of appearances. So the Futsu value searched the blog article of various appearances. We need to adjust a blog article doing the crawl and Futsu value.

5 Results and Considerations

In this paper, we proposed Kawaii Search to search blogs based on the impression formed by their text on human readers. We performed experiments to verify the utility of the Kawaii Search algorithm. By using experiments, we found that the Kawaii Value produce good results for selecting pretty. However, the Majime Value and Futsu Value did not produce good results. For our future work, we first need to correct the Majime Value and Futsu Value. Second, in addition to the six values described above, there are many factors that affect the impression of a word such as font and color. We need to expand the search algorithm by adding these types of factors. Third, when users read blog articles, the hardware used (for e.g., PC, iPad, smartphone) may have varying screen sizes, which affects the impression of the text. In this paper, we considered a standard PC or laptop screen; we need to consider the size of the screen in which the reader actually reads the blog article. Fourth, we analyzed the precision of our system based on an evaluation by 12 college students; in future, we plan to improve the search results by increasing the number of reviewers. Finally, we plan to solve these problems and improve search efficiency.

References

1. Tanabe, K.: Speech Patterns of Japanese Girls or Gals –Symbol of Identity and Opposition to Power, OPAL 3. Queen Mary, Univ. of London, London (2005)
2. Lindahl, C., Blount, E.: Weblogs: Simplifying Web Publishing. IEEE, Computer 36(11), 114–116 (2003)
3. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, NY (1983)
4. ameba, http://official.ameba.jp/
5. Kyoko, H., Katsuhiko, O.: MENKUI SEARCH: Search System Based on the First Sight of Impression Keio University, graduation thesis (2011)
6. mecab, http://mecab.sourceforge.net/
7. twitter, http://twitter.com/
8. 2channel, http://www.2ch.net/