Web Contents Evaluation based on Human Knowledge of Words

Haruhiko Takeuchi Muneo Kitajima

Institute for Human Science and Biomedical Engineering
National Institute of Advanced Industrial Science and Technology

AIST Tsukuba Central 6
Tsukuba, Ibaraki, 305-8566 JAPAN
{takeuchi, kitajima} @ni.aist.go.jp

Abstract - **The Internet has come to be used at home. Various kinds of search engines help people to find home pages which people want to read. However, we sometimes feel it difficult to understand the retrieved home pages. One of the reasons is that there exists a gap between the human's knowledge and the expected knowledge for the specific home pages. In this paper, we present a measure which shows the characteristics of home pages. For this purpose, we use a psychological word database, which was obtained by psychological experiments. We show the usefulness of the measure by applying it to sample data.**

Keywords - word, psychological experiments, content's characteristics, natural language.

I. Introduction

We acquire various kinds of information on the Internet, recently. By using search engines, such as Google, we can easily find home pages which we are looking for. However, we sometimes feel it difficult to read the specific home pages, for those are too difficult to understand the contents. Let us take an example. Suppose a person is searching a home page where there is some information for trouble shootings on the frozen computers. In a home page, there are a lot of technical terms to restart the computer, whereas in another, technical terms are not used so many. In this case, if the person is a novice user, the latter home page seems to be more suitable.

Many studies have been conducted for improving web design. But most of the quantitative methods for evaluating web sites have just focussed on page composition, page formatting, and overall page characteristics [1]. And the web content itself has not been treated well.

In this paper, we present a method to characterize the contents of home pages. There are several merits if we could know the content characteristics. For example, we can decide to read the home pages or not before reading it.

In order to deal with the web contents, we need an appropriate word database. But there is few database of this kind. The MRC psycho-linguistic database [2], where familiarity ratings were conducted for 9,054 words, is the one for English. The NTT's word database [3], where familiarity ratings were conducted for 62,356 words, is probably the only one for Japanese. Although this database reflects human knowledge of words, there are several problems for applying it to our study. Firstly, most of the subjects were restricted to 20's, and the database did not reflect the human knowledge on people of different ages. And secondly, the data, which were summed up only by men or women, were not supplied.

In our study, we make a point of user's characteristics. And we need to use the word database which reflects various subject's attributes such as age, sex, and computer experience. So we decided to start from making a new word database, and we gathered a variety of subjects. The subjective ratings were conducted for 9,050 words. Then we applied the word database to web contents evaluation. This is the first attempt to index home pages based on the psychological word database, considering their contents.

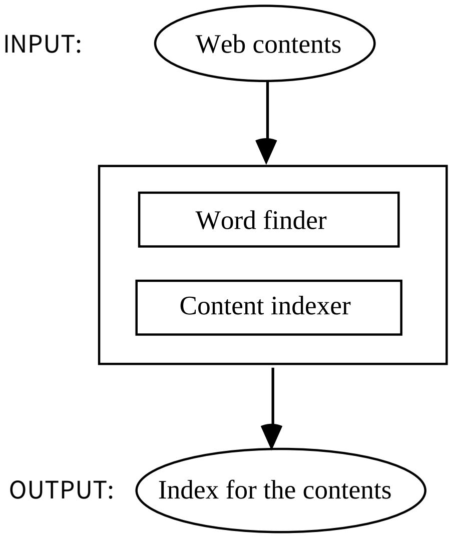

Figure 1. The Architecture of the system

II. A Measure for Indexing the Contents Features

Figure 1 gives an overview of the architecture of the system we are developing. The system can roughly be divided into two main components: a word finder component and a content indexer component. The input of the system is intended to be marked-up text documents. Plain texts can also be accepted for the input of the system. The output of the system is a numerical value between 0 and 1 for several characteristics corresponding to age, sex and computer literacy.

The word finder looks up the specific words in the contents. Here, the psychological word database is used to match the words. And the number of the times the word appears in the context is also counted. Although we can use a parser to identify the words, we preferred to use a simple algorithm, for we do not use the syntactic information now. So the words in the input texts and the words in the psychological word database are compared and the corresponding words are extracted.

The content indexer calculates the characteristics of the contents. Here, the psychological word database is also used to calculate the value for the content features. We have several options for the calculation. For example, here we will show a simple one.

Let the value of the word *i* in the database for the feature *j* be w(i, j). This value takes in between 0 to 1. We consider the features such as age, sex and computer literacy. When the word *i* is considered to be difficult to understand for those who have feature *j*, the value w(i,j) comes to be close to 1. We suppose that the word *k* (*k*=1, ... , n) is included in the input contents. Then the average word difficulty of the input content is calculated as follows:

$$f(j) = \sum_{k=1}^{n} w(k, j) / n \qquad (1)$$

By introducing a threshold level parameter p, we can identify the characteristics of the content. We compare the value of p and the value of f(j) for the features *j.* If f(j) is greater than parameter p, then we could estimate that the content is difficult for the persons who have feature *j*. When we can accept several terms which we do not know, we set the value of the parameter p to a large value such as 0.4. When we do not accept difficult words at all, we set the value of the parameter p to a small value such as 0.1 or 0.2.

III. Psychological Word Database

In order to calculate the characteristics of the home page's contents, we use a psychological word database. This word database was constructed by Novas Inc. in corporation with the authors [4]. We show the outline of this database. The main feature of this word database is that this database reflects human knowledge of words, and we can know the value of the averaged word difficulty corresponding to each feature. We can know from the database that which word is well known among the people who belong to the specific classes, e.g. male, 30's in age, and novice for computers.

Stimuli. 9,050 words were used. These words were cited from Encyclopedia of Contemporary Words [5] and e-Words [6]. Most of these words were expressed in Katakana notation. These words were mainly selected from the computer related words, and exotic words. The number of nine thousand is not enough for treating Japanese language, but if we focus just on the field of computer technology, the number would be enough.

Subjects. 57 persons joined the experiment. Their age were in between 10's to 60's. 56 person's data were

Table 1. Example from the psychological word database

(a) words which are not well known by female persons

Word	Male	Female
shock absorber	0.21	0.75
intercooler	0.25	0.75
flash pass	0.32	0.79
torque	0.11	0.57
clinch	0.07	0.54
deflagration	0.39	0.82
chronogragh	0.32	0.71
stabilizer	0.21	0.61
slugger	0.18	0.57
bowgun	0.14	0.54
impulse	0.07	0.46
Green Beret	0.04	0.43
sound card	0.32	0.68
Meteor	0.29	0.64
woofer	0.21	0.57
seajack	0.18	0.54
mustang	0.11	0.46

(b) words which are not well known by male persons

Word	Male	Female
sabrina pants	0.50	0.07
pashmina	0.75	0.29
tencel	0.71	0.25
nighty	0.68	0.21
shortning	0.54	0.07
squalane	0.82	0.32
twin knit	0.75	0.25
wrap skirt	0.68	0.18
gateau	0.57	0.07
camomile	0.82	0.29
hair wig	0.71	0.18
eye hall	0.71	0.18
petit scarf	0.68	0.11
pinky ring	0.86	0.25

used in analysis. The percentage of the subject’s age were 9% for 10's, 26% for 20's, 23% for 30's, 23% for 40's, 5% for 50's and 14% for 60's. The percentage of male and female persons was 50% each.

Method. The subjects were asked whether they understood each word. If they knew the meaning, they checked “know”. If they had read or heard the word, they checked “recognize”. If they did not know the meaning of the word at all, they checked “don’t know”. The subject’s features such as age, sex, the frequency of personal computer use, the experience years of personal computers were also recorded.

In this paper, we considered the following subject’s features: age, sex and computer literacy. As to computer literacy, those who have more than two years experience are thought to be expert users, and those who have less than one year experience are thought to be novice users. So we can classify the subjects according to these features and the subject's ratings were averaged according to these features.

We show the words which were understood differently between male and female subjects in Table 1, as an example. The number in the table shows the rate of the persons who checked “don’t know”. When the value is close to 1, most people do not know the meaning of the word. And when the value is close to 0, most people know the word well. The words listed in Table 1 (a) were not well known among female persons, and the words listed in Table 1 (b) were not well known among male persons.

IV. Application and Results

We have applied the web contents evaluation method discussed in section II to the sample data on the Internet. Here we show an example. We selected two home pages which showed troubleshootings on postscript printer errors from the Internet.

Here we will show the results corresponding to the features of computer experiences. According to the algorithm shown in Fig.1, the word finder was applied to the input contents. The words picked up in the two contents are shown in Table 2. We also show the data which

Table 2. Words in the two contents and the corresponding psychological data

(a) words used in content A

Word	Novice	Expert
postscript	0.82	0.10
memory	0.05	0.00
bug	0.41	0.10
patch	0.18	0.14
update	0.50	0.10

(b) words used in content B

Word	Novice	Expert
postscript	0.82	0.10
memory	0.05	0.00
gradation	0.18	0.05
command	0.05	0.00
object	0.27	0.00

correspond to the words in the psychological word database in Table 2.

Then the content indexer calculates the characteristics of the contents according to the equation (1). The calculated index of the content A was 0.392 for the novice feature, and 0.088 for the expert feature. The calculated index of the content B was 0.274 for the novice feature, and 0.03 for the expert feature. These results show that both of the contents are easily understood by expert users, while both of the contents are somewhat difficult to understand for novice users. If we set the threshold parameter p to 0.3, the novice feature for the content A comes to be greater than p, and it is estimated that novice computer users would feel it difficult to read the content A. And by comparing the index of the novice features for these two contents, we could estimate that the content B is more suitable for novice computer users, for the index of the content B is smaller than that of content A.

V. Conclusion

We have developed a method for characterizing the web's home pages based on the words which were used in the contents. The basic idea is to focus on the words which people can not understand. We showed the algorithm for indexing the contents characteristics. And we applied the method to a simple example, and we got a good result.

In this paper, we mainly focussed on the words in the field of computer science. For the future work, we need to develop a psychological word database which covers the specific field, in order to treat home pages in other new fields. And in order to develop a more precise algorithm for extracting the contents features, we need to conduct psychological experiments which verify the appropriateness of the measure. Although, there are several problems unsolved yet, this is the first step toward a web contents evaluation based on human psychological word databases.

Acknowledgment

The authors wish to thank Mr. H. Urokobara, Ms. J. Mine and Ms. N. Okuizumi of U'eyes novas Inc. and F. Fukuchi of Novas Inc. for considerable assistance. The authors are also grateful to Novas Inc. for providing the psychological word data for research use.

References

[1] M.Y. Ivory, R.R. Sinha, & M.A. Hearst. Empirically validated web page design metrics, *Human Factors in Computing Systems, CHI Letters*, vol. 3, No. 1, pp. 53-60, 2001.

[2] M. Coltheart. The MRC psycho-llinguistic database, *Quart. J. of Experimental Psychology*, vol.33A, pp. 497-505, 1981.

[3] S. Amano & K. Kondo. *Nihon-go no goi tokusei* (in Japanese), Sanseido, 2000.

[4] Novas Inc. Research report on human centered web design technology (in Japanese), Japan Small and Medium Enterprise Corporation, 2001.

[5] K. Horiuchi. Katakanago dictionary, *Gendai yogo no kiso chishiki (in Japanese), Encyclopedia of contemporary words*, Jiyu Kokumin-sya, 2001.

[6] e-Words. Web's page: "http://www.e-words.ne.jp/", Incept Ltd., 2001.