VISUALIZATION OF AND RETRIEVAL OF BACKGROUND INFORMATION RELATING TO WORDS IN WEB DOCUMENTS

A Visualization Interface based on Vector Representation

Kouji Shimatsuka and Tatsuhiro Yonekura
Graduate School of Science and Engeneering, Ibaraki University, Ibaraki, Japan

Keywords: Visualization, Multi Media, User Interface, Text Mining, Vector Space Model.

Abstract: When people encounter unfamiliar words, they often use tools such as search engines to obtain background information on these words. However, the semantic content of words can be complex, and it is not always possible to understand the meaning of words from textual information alone. In this paper we quantify the semantic content of words by means of a simple and convenient text-based method whereby the semantic content is constructed from linguistic, visual and auditory characteristic values. Using characteristic vectors generated in this way, users are able to visually check and search for background information on unfamiliar terms in a web document.

1 INTRODUCTION

The Web was originally used for the exchange of text-based information, but with the growth of the Internet, a vast and rich collection of multimedia information such as images, music and video content is now available online. As a result, the Web has become an extremely useful tool for looking up the meaning of words. When people encounter unfamiliar words, they often use tools such as search engines to obtain background information on these words. However, the semantic content of words can be complex, and it is not always possible to understand the meaning of words from textual information alone. In this paper we quantify the semantic content of words by means of a simple and convenient text-based method whereby the semantic content is constructed from linguistic, visual and auditory characteristic values. Using characteristic vectors generated in this way, users are able to visually check and search for background information on unfamiliar terms in a web document.

2 RELATED RESEARCHES

The field of semantic visualization methods is currently being actively researched.

Words can be related in many different ways parent-child relationships and sibling relationships can be determined in some cases, and counterfactual relationships are exhibited in other cases. However, most of these techniques only concentrate on linguistic characteristics. This is because it is generally not possible to systematically handle the weighting of words and the weighting of images and audio in a simple manner.

3 VECTOR SPACE MODEL

(Gerard Salton, Michael J. MeGill, 1983)A vector space model is a search model where a document is represented with vectors whose elements are the weightings applied to the search terms.

3.1 Weighting of Indexing Terms

3.1.1 TF-IDF

In this paper, weightings are determined by using the TF-IDF method, which is widely used for the weighting of indexing terms. The frequency at which an indexing term t_i appears in a set of documents $\boldsymbol{d}_j$ is called the term frequency (TF) and is expressed as tf_{ij}. The inverse document frequency (IDF) is used to express the specificity of these terms. The inverse document frequency *idf* is defined as follows:

Shimatsuka K. and Yonekura T.
VISUALIZATION OF AND RETRIEVAL OF BACKGROUND INFORMATION RELATING TO WORDS IN WEB DOCUMENTS - A Visualization Interface based on Vector Representation.
DOI: 10.5220/0001839204190422
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4

$$idfi = \log \frac{N}{df_i} \quad (1)$$

These two parameters are combined in order to apply weightings to the indexing terms:

$$w_{ij} = tf_{ij} * idf_j = tf_{ij} * \log \frac{N}{df_i} \quad (2)$$

3.2 Document Vector

A document vector and the entire set of documents is expressed by an m×n matrix as follows:

$$D = \begin{bmatrix} d_1 & d_2 & \cdots & d_n \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \cdots & w_{mn} \end{bmatrix} \quad (3)$$

4 PROPOSED METHOD

In this proposal, background information on words in a web document is visualized by quantifying the intended semantic content of words into three characteristic values — linguistic, visual and auditory.

To apply these weightings, the Web is used as a corpus, and MeCab is used for Japanese word segmentation and morphological analysis. The processing flow is shown in Fig. 1.

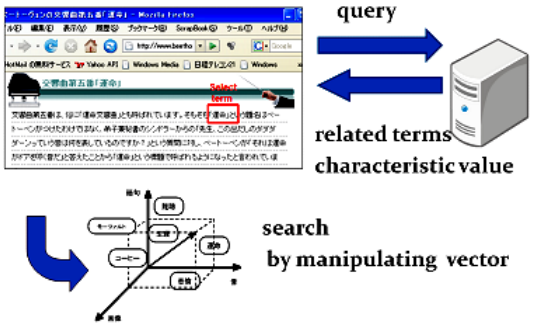

Figure 1: Flow Diagram.

4.1 Extraction from Browsing Documents

4.1.1 Extraction of Principal Terms

In the visualization of background information, the first step is to identify the subject of the Web document that the user is currently browsing. In this paper, the subject of the document is called the principal term, and the principal term is extracted as the word with the highest weighting obtained by the TF-IDF method. On the system side, the word to be searched for is transmitted along with the principal terms as a query.

4.1.2 Extraction of Related Terms

When expressing the meaning of a particular word, the existence of related terms that describe this word is an essential requirement. The greater the coverage of the related terms included in a web document, the greater the amount of information this document is likely to contain regarding these terms. The TD-IDF method is used to extract terms except for the word being searched for and the principal terms. These form a set of related terms for the query in the document bring browsed.

4.1.3 Extraction of Names

In the calculation of characteristic values, it is necessary to specify the names of image data and audio data in the web document. However, it is difficult to specify names from the content of data, and in general the name is taken from the surrounding text associated with the data. Thus for image data the name is obtained from the text of alt attributes, and for audio data the name is obtained from the text content of surrounding elements such as <a> tags and <table> tags.

4.2 Calculation of Characteristic Values

In this paper, various characteristic values are calculated from linguistic statements. It is possible to determine the degree of correspondence of words by matching them with a suitable set of related terms. It can be confirmed if characteristic value is closer to the ideal value of one page, the degree of correspondence is higher. This set of related terms is derived from the snippets appearing next to links in search engine results. The use of snippets make it possible to extract related terms that have strong co-occurrence.

4.2.1 Linguistic Characteristic Values

First, search results are obtained from a search engine based on a query. Next, snippets consisting of the summarized search results are analyzed to calculate the TF-IDF values of each word. The 10

highest ranking terms (other than the search keywords) that are obtained in this way are assumed to be related terms that are suitable for the query. This set of related terms is treated as a query vector, and its inner product with the document vector is calculated with a weighting of 1. The value calculated by the inner product is the sum of the weightings of the related terms in the document being browsed. The total weighting of the related terms in the document the user is looking at is defined as *Total*, *Total* is calculated as follows:

$$Total = \mathbf{d_j} \cdot \mathbf{q} = \sum_{i=1}^{m} w_{ij} q_i \quad (4)$$

Finally, the ideal of one page is calculated.

The entire set of documents from search results is defined as ***R***, the total weighting of related terms in the search results is defined as *RTotal*, analyzed number of pages is defined as *tn*, the ideal value of one page is defined as *OneTotal*.

$$\mathbf{R} = \begin{bmatrix} r_1 & r_2 & \cdots & r_{1tn} \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1tn} \\ r_{21} & r_{22} & \cdots & r_{2tn} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mtn} \end{bmatrix} \quad (5)$$

$$OneTotal = \frac{RTotal}{tn} = \frac{\mathbf{R} \cdot \mathbf{q}}{tn} = \frac{1}{tn} \sum_{j=1}^{tn} \sum_{i=1}^{m} r_{ij} q_i \quad (6)$$

Then the characteristic value is calculated as follows:

$$\frac{Total}{OneTotal} \quad (7)$$

4.2.2 Visual Characteristic Values

Images in web pages are often described in some way by the page's text content. Since this descriptive text often includes terms that are related to the image's visual appearance, the visual appearance information provided by these words can be used to apply weighting to the content expressed by the image. As in the calculation of linguistic characteristic values, suitable related terms relating to the visual appearance information are acquired from the snippets. The discrimination of words expressing visual appearance information is performed by using a dictionary in which the visual appearance attributes have been previously declared.

The visual characteristic value is also calculated as follows:

$$\frac{imgTotal}{imtOneTotal} \quad (8)$$

A list of visual appearance attributes for the Golden Pavilion temple in Kyoto is shown in Table 1(browsing document 'http://onseno06.com/').

Table 1: List of visual appearance attributes for the Golden Pavilion temple "Kinkakuji-ji(金閣寺)".

term	browsing document	ideal weight of one page(*in*=5)
山荘	8	9
庭園	0	8
池	5	8
三層	0	7
二層	0	5
金箔	7	5
境内	6	4

4.2.3 Auditory Characteristic Values

Music and speech data is also often associated with text descriptions in web documents including information such as the track name or onomatopoeia words. Auditory information obtained from these words can be used as weightings representing the audio content. As in the calculation of linguistic characteristic values, suitable related terms relating to the audio information are obtained from snippets. The discrimination of words expressing audio information is performed using a dictionary in which the audio attributes have been previously declared. The the auditory characteristic value is also calculated as follows:

$$\frac{sndTotal}{sndOneTotal} \quad (9)$$

4.3 Creation of Characteristic Vector

From the calculated characteristic values, a characteristic vector is created for the word to be searched. If v1, v2 and v3 correspond to the linguistic, visual and auditory characteristics, then the characteristic vector of the word to be searched can be expressed as follows:

$$\mathbf{v} = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \quad (10)$$

4.4 Search Background Information

In case Linguistic characteristic values, in order to

make the value of formula (11) and search results weight more than 0 for every *i*, search results is obtained.

$$w_{ij} - \frac{1}{tn}\sum_{j=1}^{tn} r_{ij} q_i \qquad (11)$$

5 EXPERIMENTAL VISUALIZATION/ SEARCH INTERFACE

The characteristic vectors of the words and the related terms were mapped into three dimensions, and by manipulating the characteristic vectors and text strings, it was possible to obtain a specified quantity of information from multiple types of media at the same time, and to check unfamiliar words from the emphasized related terms. If the related terms are musical terms then they are mapped closer to the auditory axis, and if they are terms expressing visual appearance information then they are mapped closer to the visual axis, whereby different terms are mapped closer to their related axes.

6 CONCLUSIONS

In this proposal we have described how the semantic content of words can be quantified by defining the semantic content intended by words as being expressed by three types of characteristic quantities — linguistic, visual and auditory. The characteristic vector generated in this way allows users to visually check and search for unfamiliar words. In this way, it should be possible for users to efficiently ascertain the semantic content intended by unfamiliar words.

Figure 2: Vector manipulation.

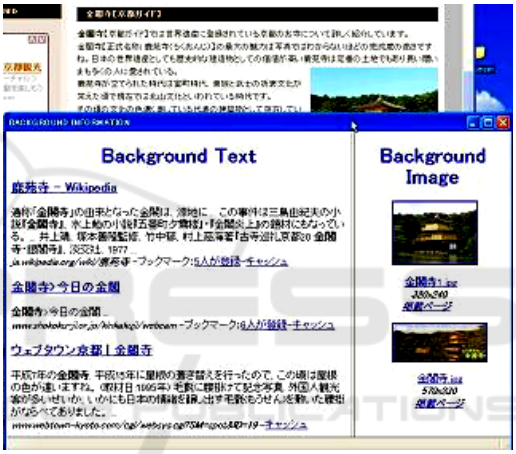

Figure 3: Acquisition of search results.

ACKNOWLEDGEMENTS

This work was partially supported by the JSPS Grant-In-Aid no.18300027.

REFERENCES

Gerard Salton, Michael J. MeGill, 1983. Introduction to Modern Information Retrieval. MeGraw-Hill.

E. Chisholm, T. Kolda, 1999. New term weighting formulas for the vector space method in information retrieval. Technical Memorandum ORNL-13756.

K. Kita, K. Tsuda, and M. Shishibori, 2002. Information Retieval Algorithms. Kyoritsu Shuppan Press.

Kanada, Y., 1999. A Method of Geographical Name Extraction from Japanese Text for Thematic Geographical Search. 18th International Conference on Information and Knowledge Management (CIKM'99), pp. 46-54.

Matsuo, Y., Ishizuka, M., 2002. Keyword Extraction from a Document using Word Co-occurrence Statistical Information.Journal of the Japanese Society for Artificial Intelligence(17-3D), pp.217-223.