A System to Construct an Interest Model of User Based on Information in Browsed Web Page by User

Kosuke Kawazu[1], Masakazu Murao[2], Takeru Ohta[3], Masayoshi Mase[3], and Takashi Maeno[2]

[1] Keio University, Graduate School of System Design and Management, 4-1-1 Hiyoshi, Kouhoku, Yokohama, Kanagawa 223-8526, Japan
[2] Keio University, Graduate School of System Design and Management, Japan
[3] Keywalker. Inc, 6th floor Sprit Buildin 3-19-13 Toranomon, Minatoku, Tokyo, 105-0001, Japan
do-my-best@a6.keio.jp

Abstract. In these days, they expect that computers comprehend characteristics of the user, for example interest and liking, to interact with computers. In this study, we constructed a system to construct an interest model of the user based on information in browsed Web pages by the user by extracting words and interword relationships. In this model, metadata is appended to words and interword relationships. Kinds of metadata of words are six, personal name, corporate name, site name, name of commodity, product name and location name. And metadata of interword relationships is prepared to clarify relationships of these words. This system makes a map by visualizing this model. And this system has functions to zoom and modify this map. We showed efficacy of this system by using evaluation experiment.

1 Introduction

In recent years, we became possible to do a lot of works by using computers with the high functionality in computers. As the result, it becomes difficult that computers comprehend the user requirements. Today it is noticed to model characteristics of the user, for example interests and likings, to comprehend the user requirements through interactions between the user and computers.

It was formerly conducted that some studies to model characteristics of the user. There are two methods to model the user. One method is collaborative filtering to model characteristics of the user by using information of a lot of users. Another method is information filtering to model characteristics of a user by using information of a user [1]. Collaborative filtering is a method to extract action patterns of user communities [2]-[5]. So, detailed characteristics of the user can't be model by using collaborative filtering. In this study, information filtering is chosen to model detailed characteristics of the user.

There are two methods in information filtering. One method is explicit one. This is that interested fields and keywords are evinced by the user. Another method is implicit one. This is that computers extract automatically interested keywords of the user by using

J.A. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2009, LNCS 5612, pp. 304–313, 2009.

actions at the time of browsing [6]. Explicit method is highly-loaded because the user must evince analysis objects every time. And model of characteristics of the user is limited because of limited analysis objects [7]. In this study, implicit method is chosen.

In the study using implicit method, there are Corin's study and Murata's study. Corin constructed a personalized start page by using Web access history [8]. Murata established Site-Keyword graph to visualize and extract interest of the user by using logdataes [9]. Site-Keyword graph is a graph which has titles of Web pages and search keywords as vertices and time ordering as sides. Thus, heretofore, vertices are titles of Web pages and search keywords. And sides are time ordering and occurrence rate in existing models. But if vertices are titles of Web pages and search keywords, characteristics of the user aren't clear. And if sides are time ordering or occurrence rate, it is difficult to conjecture objects of interest of the user because the user's actions are inconsequence.

So purpose of this study is to establish a system to construct an interest model of the user which has two features by using information in browsed Web pages by the user:

- Detailed model of the user; and
- Easy model of the user to use.

2 Concept Design

In this chapter, in order to fulfill the demands stated in the previous chapter, concept design is clarified.

2.1 An Analysis Object for Making the Model

As stated in previous chapter, proposed system is that builds the detailed model of the user to present information for which the user hopes. It is necessary to consider an analysis object for making the detailed model of the user. In this study, the detailed interest model of the user is defined as clarifying the interword relationship by using each word as each vertex. Hence, in this study, analysis object is defined as sentences of browsed Web page by the user. This system is able to clarify the object of user's interest by extracting object of user's interest word by word by using sentences of Web page. That will be effective to present information for which the user hopes. Concretely, proposed system gets words and interword relationships by extracting title and body text of browsed Web page by the user.

2.2 Addition of Metadata

As stated in previous chapter, proposed system is that builds the user's interest model that is used easily to present information for which the user hopes. In this study, easy model of the user to use is defined as being able to arrange and categorize object of user's interest easily. As a result, proposed system will be able to conjecture user's requirement by using the model. So, it is effective to arrange and categorize words and interword relationships. Hence, proposed system adds metadata into each word and each interword relationship to arrange and categorize them.

Table 1. Example of metadata of word

Name of Metadata	Example
Personal name	George Washington Marilyn Monroe
Corporate name	SONY NASA
Site name	Yahoo Google
Name of commodity	Xbox360 Freelander

Table 2. Example of metadata of interword relationship

Interword Relationship	Name of Relation
Personal name – Personal name	• Collaboration
Personal name – Corporate name	• Member
Personal name – Name of commodity	• Connected commodity
Corporate name – Corporate name	• Alliance • Competitive • Connected corporation • Same area • Group
Corporate name – Name of commodity	• Development, Sale
Site name — Name of commodity	• Connected site

Metadata of Words

The metadata of words is clarified. As stated in previous, it is necessary to add leading concept that this word is corporate name into a word such as Google to arrange and categorize words. The ease of use of interest model of the user will increase by adding these leading concepts as metadata into words.

In this world, there are various words. Therefore, there are various metadata that is the leading concept of them. It is difficult to find out all kinds of metadata and make proposed system get them. Then, we categorized Web search words that were released by various internet search sites [10]-[13]. As a result, it was found that the Web search words used frequently by user are:

- Personal name such as George Washington and Marilyn Monroe.
- Corporate name such as SONY and NASA.
- Site name such as Google and Yahoo.
- Name of commodity such as Xbox360 and Freelander.

Therefore, in this system, metadata of words is constructed from personal name, corporate name, site name and name of commodity. Proposed system extracts words that correspond to metadata and adds the metadata into the words. Table 1 shows examples of words that can be obtained and the kinds of metadata. These words can be obtained by analyzing sentences of browsed Web page by the user.

Metadata of Interword Relationships

The metadata of interword relationships is defined. Previously, the interword relationship was defined by using occurrence rate of word that is represented by Takashiro's research [14]. But the method defined by using only occurrence rates of a word can not clarify difference of interword relationships between words. It is effective to define the relationships by using not only occurrence rate of word but also metadata to clarify them to arrange and categorize the model. Interword relationships are decided by using semantic analysis based on situational context in sentences of browsed Web page by the user. Hence, proposed method defines interword relationships by using two kinds of analysis based on occurrence rate of word and semantic relation of word at the same time. In this study, words proposed system extracts is defined by four kinds of metadata that are personal name, corporate name, site name and name of commodity. Hence, it is necessary to define interword relationships in order to connect these words. There are various relationships between the words. Therefore, it is difficult to define all of them. Hence, the relationships that are specified and extracted easily in sentences of Web page are defined. Table 2 shows example of interword relationships that can be obtained and kinds of metadata.

2.3 Outline of Proposed System

Outline of proposed system based on methods stated above is clarified. The function of proposed system is divided into two functions. The first function extracts words and interword relationships, adds metadata into each of them, and constructs interest model of user by analyzing sentences of browsed Web page by user. The another function visualizes interest model of user for user to browse own model. As a result, user will be able to modify own model properly.

Fig. 1 shows outline of processes of proposed system.First, user visit a Web page. As a result, this system obtains user's ID, time, the Web page of URL and sentences in the website. Second, this system analyses sentences of the Web page. As a result,

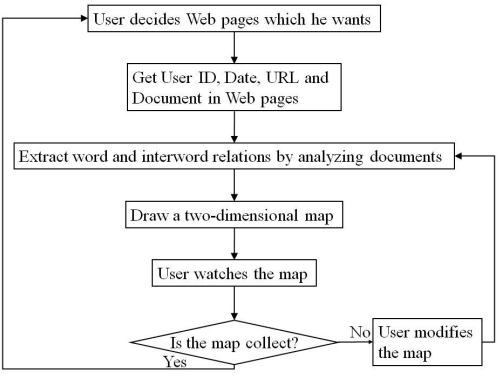

Fig. 1. Outline of Proposed system

the sentences are decomposed to words. And, the system adds metadata into each word and interword relationship. Third, the system constructs map by arranging these words and interword relationships in two dimension spaces as vertices and sides. Finally, this map is visualized for user to modify it.

3 Detail Design and Building

This system is constructed by two subsystems. One subsystem is a database. Another subsystem is an user interface. In this chapter, algorithms of this database and this user interface are described.

3.1 Algorithm of Database

In this study, URL and document of Web pages are extracted by installing plug-in of our own composition on Firefox. In this database, vertices, layout of vertices and sides to make a map are defined by analyzing information of Web pages browsed by the user. In this database, it processes by the following flows:

1. URL of a browsed Web page by the user is received from Firefox.
2. HTML is received.
3. Documents are extracted.
4. Situation Analyze(SA) is performed to the acquired documents.
5. A map of the user is made in two-dimensional space.

SA has two features:

- Metadata of words is appended by performing morphological analysis and syntax analysis to the acquired documents; and
- Metadata of interword relationships is appended by using semantic network and analyzing documents of browsed Web pages by the user.

In the previous chapter, we described that words which were appended metadata of personal name, corporate name, site name and name of commodity are extracted. A rule-based database is established to extract these words. This is a rule to define metadata, personal name and corporate name etc, from the acquired documents. Fig. 2 shows a flow between input to output. Words and metadata of words are outputted from the acquired documents by checking the rule in the case of coinciding.

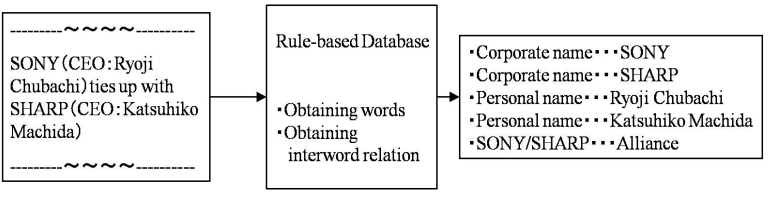

Fig. 2. Flow between input to output

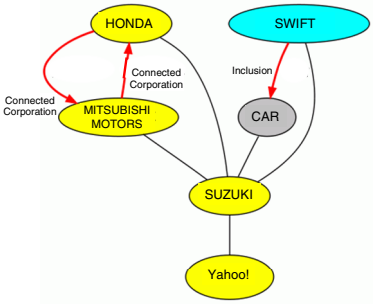

Fig. 3. Visualized map

Table 3. Color of Metadata

Name of Metadata	Color
Personal Name	Green
Corporate Name	Yellow
Site Name	Red
Name of Commodity	Blue
Product Name	Gray
Location Name	White

In this study, interword relationships are defined by combining with two algorithms. One algorithm is interword relationships based on occurrence rates of a word. Interword relationships based on occurrence rates are defined by weights of interword relationships. Weights W of interword relationships are calculated by:

$$W_{W1\&W2} = \log(N_{W1\&W2}) \times \log\left(1 + \frac{N_{W1\&W2} N_{URL}}{N_{W1} N_{W2}}\right) \quad (1)$$

$W_{w1\&W2}$ is weight of relationships of the word2 to word 1. N_{URL} is total number of browed Web pages by the user. $N_{W1\&W2}$ is the number of Web pages in which word 1 and word 2 are contained at the same time in browsed Web pages by the user. N_{W1} is the number of Web pages in which word 1 is contained in browsed Web pages by the user. N_{W2} is the number of Web pages in which word 2 is contained in browsed Web pages by the user. Words which are acquired by SA are calculated by using this calculating formula. And average weight of all the words that have relationship in word 1 and standard deviation are calculated. Then, words which have weight of more than the sum total of average weight and standard deviation are regarded as having relationships.

Another algorithm is interword relationships based on meaning in documents. interword relationships based on meaning are appended on relationships based on occurrence rates. A rule-based database is established to extract interword relationships based on meaning as well as acquisition of words. Metadata of interword relationships is appended by checking the rule and documents of Web pages in the case of extracting relationships, for example connected commodity.

Finally, layout of vertices is defined by using Graphviz which is software. As the result, vertices are arranged in two-dimensional space in a single layer. And then, vertices have x-coordinate and y-coordinate. This information is outputted as XML.

3.2 Algorithm of User Interface

In the user interface, worked information in the database is acquired. And a user can interact with the computer by using this information. In particular, firstly, contents of vertices, coordinates of vertices and interword relationships are defined by receiving in Flash information which was outputted as XML. Secondly, the computer makes a

map by defining shapes of vertices, colors of vertices and sides. Finally, the user can interact with the computer by modifying his model. The map has functions to modifying the model of the user, for example a function to zoom and his map, a function to move vertices of his map. Fig. 3 shows a sample of the map which is made actually. And Talble 3 shows relationships between metadata of words and color. In this sample map, vertex of Yahoo draws by yellow because site name is sometimes recognized faultily as corporate name.

4 Verification Experiment

In this chapter, the two experiments were conducted to verify effectiveness of proposed system.

4.1 Verification of Proposed System

The first experiment was conducted by two procedures to verify validity of kinds of metadata and extraction of words. First, examinees browsed the Web pages in thirty minutes. The second task for examinees was to extract the words he has been interested in from the Web page he browsed and to write down the words and interword relationships freely. What examinees wrote down was constructed from the words as vertex and the relationships as sides. This experiment was conducted by eight examinees in their twenties and thirties.

Verification of Kinds of Metadata of Words
In this study, the metadata of words was constructed from personal name, corporate name, site name and name of commodity. Therefore, it is necessary to verify content rate of these four kinds of metadata. Content rate of metadata of words, $R_{inclusion,}$, where N_{ALL} and N_{TAG} are the total number of word examinees wrote down and the number of word classified into the four kinds of metadata, respectively, is expressed by

$$R_{inclusion} = \frac{N_{TAG}}{N_{ALL}} \tag{2}$$

Fig. 4 shows the relationship between content rate of metadata of words and the number of kind of metadata. As shown in Fig. 4, proposed system constructed from four kinds of metadata can express only 37.4% of words in which examinees was interested. Consequently, two kinds of metadata of words that are location name and product name were added into proposed system. At the same time, two kinds of metadata of interword relationships that are inclusion and location were added into proposed system. As a result, as shown in Fig. 4, the system constructed from six kinds of metadata can express 57.1% of words in which examinees was interested.

Verification of Words Extracted by Proposed System
The words extracted by proposed system were verified by comparison with the words examinees wrote down. Concordance rate of words, R_{match}, where N_{SYS} are the number

of the words extracted by proposed system that coincide with the words examinees wrote down, is expressed by

$$R_{match} = \frac{N_{SYS}}{N_{TAG}} \tag{3}$$

Fig. 5 shows the relationship between concordance rate of words and the number of kind of metadata. As shown in Fig.5, in case of four kinds of metadata, concordance rate was 69.9% and in case of six kinds of metadata, concordance rate was 75.2%.

4.2 Verification of Effectiveness of Proposed System

The sensory evaluation experiment were conducted to verify effectiveness of proposed system that construct the user's interest model by using metadata of words and interword relationships. First, examinees browsed Web pages freely. At the same time, proposed system construct the interest model of examinee's. In this experiment, the task for examinees was to compare the two maps. The one is the model didn't include metadata, the other is the map included it. The map that didn't have metadata shows words without changing the color of vertices by kinds of word such as personal name or corporate name. And, the map shows interword relationships without writing the semantic of relationship by kinds of interword relationships. As a result, examinees didn't understand what kinds of relationship were. In contrast, as shown in Fig. 3, the map that has metadata shows words that have the color of vertices and interword relationships that have the semantic of them. These two models were compared by examinees in sensory evaluation experiment. The five evaluation items were:

(a) Easiness of seeing this map.
(b) Degree of similarity between the objects of your interest and this map.
(c) Degree of contain the keywords you were interested in.
(d) Amount of extra words in this map.
(e) Adequacy of interword relationships.

This sensory evaluation experiment was conducted by eight examinees. Fig. 6 shows averages and standard deviations of the evaluation of each evaluation item. As a result

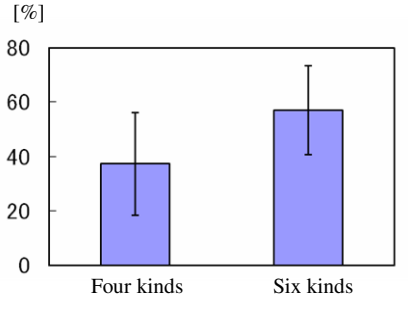

Fig. 4. Content rate of metadata of words

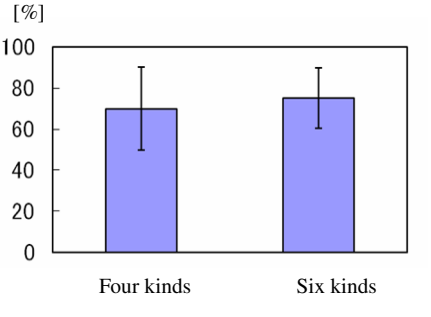

Fig. 5. Concordance rate

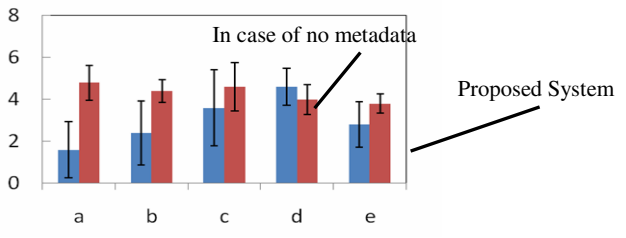

Fig. 6. Result of sensory evaluation

of two-sample t-test, evaluation item of (a) showed significant differences of 1% of significant level and evaluation item of (b) and (e) showed differences of 5% of significant level.

5 Consideration and Future Task

Fig. 5 shows that the system constructed from six kinds of metadata can express 57.1% of words in which examinees was interested. In the future, it is effective to add some kinds of metadata in order to model user's interest more adequately. However, extra extracted words in which users are not interested would increase by increasing the kind of metadata. To solve this problem, it is necessary to propose algorithm for distinguishing between the words in which the user is interested and the words in which the user isn't interested. The development of this algorithm and the increase of $R_{inclusion}$ would be attempted in the future.

The words extracted by proposed system were verified by using R_{match}. Fig. 6 shows that there is no significant difference between concordance rate of words of four kinds of metadata and that of six kinds of metadata. This means that concordance rates of words about each kind of metadata are almost constant and high level. In the future, the increase of concordance rate of words would be attempted.

As shown in Fig. 6, it was found that the easiness of seeing the user's interest model increase by using metadata of words and interword relationships. However, it was found too that examinees sometimes feel that metadata of interword relationships is not appropriate. It will be the future task to solve this problem by adding the kinds of metadata of interword relationship appropriately.

6 Conclusion

In this study, we constructed a system to construct an interested model of the user based on information in browsed Web pages by the user. This system is a foundation to present appropriate information to a user. This system can model 57% of interest objects of the user by using personal name, corporate name, site name, name of commodity, product name and location name. This became clear from evaluation experiment. And we constructed easy model to use for the user by appending metadata to words and interword relationships. In the future, we will construct a system to present appropriate information to a user by applying this system.

References

1. Huangr, H., Fujii, A., Ishikawa, T.: The individualized technique of the information retrieval based on the Web community. In: The Association for Natural Language Processing Annual Conference, vol. 11, pp. 1006–1009 (2005)
2. Niwa, S., Doi, T., Honiden, S.: Web Page Recommender System Based on Folksonomy Mining. Transactions of information Processing Society of Japan 47(5), 1382–1392 (2006)
3. Kazienko, P., Kiewra, M.: Integration of relational databases and Web site content for product and page recommendation. In: Database Engineering and Applications Symposium, IDEAAS (2004)
4. Golovin, N., Rahm, E.: Reinforcement Learning Architecture for Web Recommendation. In: Proceedings of the International Conference on Information (2004)
5. Claypool, M., Gokhale, A., Miranda, T., et al.: Combining Content-Based and Collaborative Filters in an Online. In: Proc. ACM SIGIR 1999 Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, California (1999)
6. Hijikata, Y.: User Profiling Technique for Information Recommendation and Information Filtering. Journal of Japanese Society for Artificial Intelligence 15(5), 489–497 (2003)
7. Kantor, P.B., Boros, E., Melamed, B.: Capturing Human Intelligence in the Net. Communications of the ACM 43(8), 112–115 (2000)
8. Anderson, C.R., Horvitz, E.: Web Montage: A Dynamic Personalized Start Page. In: Proceedings of the 11th World Wide Web Conference (WWW's 2002) (2002)
9. Murata, T., Saito, K.: Extraction and Visualization of Web Users' Interests Using Site-Keyword Graphs. Journal of Japan Society for Fuzzy Theory and Intelligent informatics 18(5), 701–710 (2006)
10. http://www.google.com/press/zeitgeist2005.html
11. http://searchranking.yahoo.co.jp/ranking2008/general.html
12. http://searchranking.yahoo.co.jp/ranking2008/general.html
13. http://www.technorati.jp/ranking2006/
14. Takashiro, T., Takeda, H.: Acquisition and Organaization of Personal Knowledge through WWW Browsing. Institute of Electronics, Information, and Communication Engineers J85-D-1(6), 549–559 (2002)