Web Information Retrieval Using Ontology for Children Based on their Lifestyles

Mika Nakaoka
Graduate School of Informatics
Kyoto University
Yoshida-Honmachi, Kyoto, Japan
nakaoka@dl.kuis.kyoto-u.ac.jp

Yukari Shirota
Faculty of Ecnomics, Gakushuin University
1–5–1 Mejiro, Toshima-ku, Tokyo, Japan
yukari.shirota@gakushuin.ac.jp

Katsumi Tanaka
Graduate School of Informatics
Kyoto University
Yoshida-Honmachi, Kyoto, Japan
tanaka@dl.kuis.kyoto-u.ac.jp

Abstract

Users of existing Web information retrieval systems need to input proper keywords through which the correct retrieval data can be found. It is, however, too difficult for children to identify appropriate keywords due to their lack of ability in generalizing concepts describing what they would like to search for. Therefore, a great deal of irrelevant data is included in the results. In the paper, we propose a "Kid's Lifestyle Ontology" so that the system can infer what they would like to search for to resolve ambiguities. The retrieval system can identify appropriate keywords that have strong relationships with concepts common to their lifestyles, and then the related search keywords can be ranked above other keywords. This is based on the episodic memories of children that help them discover other keywords. As a result, children can refine their queries gradually. In the paper, we also propose a method of constructing the ontology incrementally and recursively.

1. Introduction

Today's technologies are changing the way children learn and play. With innovative tools, children can read books, play games, and communicate with computers and others. Today's researchers developing new technologies for children must strive to understand their unique needs. Our research goal was to look for easy ways for children to access electronic information, especially the retrieval of Web information. There are currently many elementary schools trying to use Web information retrieval systems to support pupils' theme-based "investigative learning." As a new innovative tool to support that, Sumiyoshi proposes an educational broadcasting service that supports not only passive learning through the simple watching of broadcast programs but also active learning where students search for and collect programs regarding their own purposes, such as making presentation materials and exchanging opinions by using agent technologies.[1] To make the theme-based "investigative learning" currently taught by many Japanese elementary schools to be effective, new retrieval methods for children are required.

In existing Web information retrieval systems, users need to input proper keywords through which the correct resulting retrieval data can be found. It is, however, too difficult for children to identify appropriate keywords due to their lack of ability in generalizing concepts describing what they would like to search for. Therefore, a great deal of irrelevant data is included in the results. In this paper, we propose a "Kid's Lifestyle Ontology" so that the system can infer what they would like to search for to resolve ambiguities. The retrieval system can identify appropriate keywords that have strong relationships with an ontology (a shared conceptualization) in their lifestyles, and then the strongly related search keywords can be ranked above other keywords. The ontology can be generated based on the episodic memories of children that help them discover other keywords. As a result, children can refine their queries gradually.

We propose the Kid's Lifestyle Ontology for children's Web retrieval tools in the next section. The Web retrieval process using the ontology is described in Section 3. The

Proceedings of the 21st International Conference on Data Engineering (ICDE '05)
1084-4627/05 $20.00 © 2005 IEEE

IEEE COMPUTER SOCIETY

episodic memory creation process is explained in Section 4. The paper ends with a discussion and our conclusions.

2. Web Information System for Children

Our research goal was to develop an innovative tool to help preschool search Web pages. The way preschool children (ages 4–6) remember things is that their recall is related to their episodic memories [2], i.e., they have not yet developed capabilities for abstraction and generalization. For children to remember something, they must first find when and where the episode happened, and then identify and remember what it was. When they are searching for something with Web information retrieval systems, they need to use episodic memories related to what they want to search for. In particular, the most important parts of their memories are memorized in images. Therefore, episodic memories and images can be used to support children in their Web retrieval tasks.

In general, user profile information can be used to retrieve Web pages matching the user's interests. Existing Web commerce applications named shopbots generate and use a user profile asking the user for his/her preferences.[3, 4, 5] Children's lifestyle ontologies are more helpful in searching for Web documents, compared with their user profiles because they cannot describe their preferences to the systems. Ontologies have been introduced to facilitate knowledge sharing and reuse between various agents, regardless of whether they are humans or artificial in nature.[6]

We thus propose an ontology for children called the "Kid's Lifestyle Ontology" to define children's lifestyles that allows information processing by computer.

The proposed ontology consists of their daily life events and special season events such as Christmas. The class hierarchy for the ontology can be dynamically and gradually created and modified by using the current information on the Web pages, especially Weblog pages. Parenting diaries on Weblogs by mothers and nursery school teachers, Miki House's Web pages on child nursing, and Benesse's Web pages on child nursing, are especially helpful in collecting the latest information about children's lifestyles and episodes. For example, popular Christmas presents for children, children's infectious diseases, good guides for picture books, and parent's guides to primary school, are written on Weblog pages, reflecting the children's current activities and episodes.

The problem with the existing ontology is that it does not reflect the latest information because it is static and fixed. Our proposed "Kid's Lifestyle Ontology", on the other hand, can include the latest information retrieved from the Web and can be gradually modified and improved in quality. Let us consider characters in popular television programs and animation. New heroes continuously appear. For example, in "pocket monsters"• •which is a popular television cartoon program, a new type of monster appears in each episode. The EPG (Electronic Program Guide) to Web pages can be used to find such new information on characters.

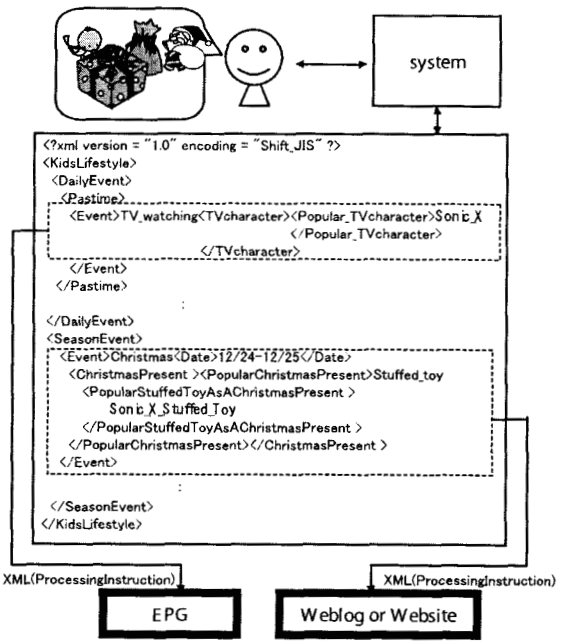

Figure 1. Outline of Kid's Lifestyle Ontology

The "Kid's Lifestyle Ontology" consists of event titles and related information that explains the details of the events. The retrieved episodes that occur in children's daily lives are stored as subclass information of the event class. For example, episodes related to "Christmas" are "Christmas presents are placed around the Christmas tree," "Santa Claus brings a present during the night before Christmas," and "We have a Christmas pudding and a Christmas cake,". From the episodes, keywords are selected and added as the subclass of the event "Christmas." The episodes also include concrete 5W1H information. Therefore, by asking the child when and where the episode happened, the system can infer which kind of episode happened to the child. Such questions asked of the child can help him or her to discover other keywords and refine his/her queries gradually.

3. Web Retrieval Process Using Ontology

A Web retrieval process using the "Kid's Lifestyle Ontology" is presented in this section. How to generate and modify the ontology as a back office process is also ex-

IEEE COMPUTER SOCIETY

plained (See Fig.1). In this example, suppose that Harry aged six wants to find a stuffed toy of a cartoon character called "Sonic" that his friend received as a Christmas present from his parents. Harry wants to know about the stuffed toy. The following are interactions between the system agent and Harry:

(1) Harry inputs "SONIC" as a search keyword. The resulting data include many irrelevant pages such as participating Sony dealers.

(2) Agent: "Where did you see SONIC?"

(3) Harry inputs "TV."

(4) The system dynamically collects a set of character names in popular television programs from EPG pages on the Web, issuing the query "What is a popular children's television character?". The resulting data are attached to the ontology as related information to the event "television watching." Suppose that "Sonic" exists in the resulting data.

(5) Agent: "You are searching for Sonic X that you can see at eight o'clock on Sundays on television."

(6) Harry: inputs "Yes." Based on Harry's answer, the system is convinced that the target is "Sonic X."

(7) Agent: "What would you like to know about "Sonic X?"

(8) Harry inputs "stuffed toy." The system finds the keyword "stuffed toy" under the Christmas event in the "Kid's Lifestyle Ontology." The subclass hierarchy consists of "Christmas," "Christmas presents," "popular Christmas presents," and "popular stuffed toys as a Christmas present." Then, the system dynamically collects the latest information about the stuffed toy, issuing the query "What is a popular stuffed toy given as a Christmas present?" to the parenting Weblog pages. The resulting data are attached as information related to "popular stuffed toys as a Christmas present." Suppose that the resulting data includes "Sonic X stuffed toy."

(9) The agent shows a retrieved photo of the stuffed toy that was embedded on the Web pages. After this, a great deal of data about Sonic X can be retrieved and given to Harry.

During the process, the dynamically retrieved data are attached as part of the ontology in the form of XML tags (see Fig.1). The queries issued by the system to find new information can be implemented as a computer application module. For example, XML PI (Processing Instructions), which is a procedural element, can be used: An XML processor will pass the PI through to the application.

4. Episode Memory Creation Process

In this section, we discuss episodic memory creation methods. First, we will report experiment results obtained by extracting episodic memories from Weblog pages. The aim of this experiment was to extract keywords concerning Christmas from parenting Weblog pages written by mothers. The target Weblog pages were selected in advance through human tasks. We used existing algorithms such as "WWW/Blog/Identify" [7] for the Weblog assessment module.

We selected words that appeared the most frequently as keywords in the experiment. As a result, the extracted keywords were "present," "children," "Christmas," and "Santa Clause," as shown in Fig.2. We used ChaSen[8] as the Japanese morphological analysis system.

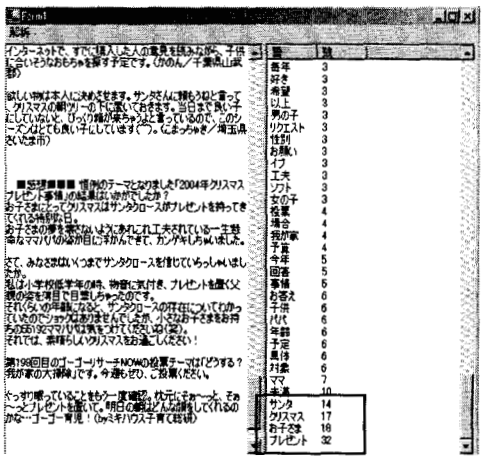

Figure 2. Results of relations for Christmas

The extracted keywords can be candidates for the event subclasses of the "Kid's Lifestyle Ontology." To determine whether the keyword is appropriate as the subclass title, some heuristic screening algorithms are required. Algorithm development will be one of our next research tasks. The episodic memory creation process is executed as follows:

(1) The system creates the subclass of the event for which title is the extracted hot keyword.

(2) The system collects a set of episodic memory sentences that include the extracted hot keyword as the subject or the object.

This creative process is repeated dynamically and recursively.

Let us next consider existing work related to extracting keywords from Weblogs.

The bursty word selection algorithms by Kleinberg[9] may be helpful in extracting the latest keywords from Weblogs. The 'burst of a word' is a phrase that indicates a drastic increase in frequency as the topic begins to emerge.

Kleinberg's original algorithm to identify bursts has been extended by Okumura et al.[10] In Okumura's blogWatcher, which is a Japanese Weblog monitoring system, subjective (evaluative) expressions in blog entries are automatically annotated by using a naive pattern matching algorithm with a manually constructed dictionary of evaluative ex-

IEEE COMPUTER SOCIETY

pressions. In blogWatcher, the extracted information is only noun phrases. In Kurashima's research, on the other hand, a Japanese sentence could be extracted by using a regular expression pattern template that included a postpositional particle and a noun phrase.[11] Similar research was done by Tezuka focusing on the co-occurrence of place names and postpositional particles.[12]

Some practical freeware for Japanese natural language processing has currently become available.[13] For example, the POSUM freeware system extracts important sentences using the freeware, Lexical Chainers. Not only noun phrases but also episode sentences have to be extracted to develop the "Kid's Lifestyle Ontology". In future work, we intend to find a suitable algorithm to extract episodic sentences from Weblogs.

5. Discussion and Conclusion

We proposed the "Kid's Lifestyle Ontology" to help children retrieve Web information. Its featured (1) daily and season events that happened in their lifestyles, and (2) could be dynamically and gradually created and expanded by collecting the latest information from Web pages, especially that from blogs written by their mothers and nursery school teachers. This ontology helped children find information they required from the Internet, because a great deal of their typical lifestyle information is included there.

When children want to search for something, they use episodic memories and images to express what they want. To identify their target object, place information on "where the event happened" is useful. Thus, event information in the ontology has to include "where" information and some agent systems are required to ask children location information. When developing methods of extracting children's episodic memories, we have to strive to understand their unique and self-centered ways of expressing "where" information. Through extraction methods, implementation, and empirical studies on episodic memories, we expect to be able to better understand young children's unique ways of remembering and expressing things.

Acknowledgement

This work was supported in part by the Japanese Ministry of Education, Culture, Sports, Science and Technology under a Grant-in-Aid for Software Technologies for Search and Integration across Heterogeneous-Media Archives, a Special Research Area Grant-In-Aid For Scientific Research (2) for the year 2004 under a project titled Research for New Search Service Methods Based on the Web's Semantic Structure (Project No, 16016247; Representative, Katsumi Tanaka), and the Informatics Research Center for Development of Knowledge Society Infrastructure (COE program by Japan's Ministry of Education, Culture, Sports, Science and Technology).

References

[1] H. Sumiyoshi, I. Yamada, Y. Murasaki, Y.B. Kim, N. Yagi, and M. Shibata, "Agent Search System for A New Interactive Education Broadcast Service", NHK STRL R&D No.84, Mar. 2004

[2] Linton ,M : Transformations of memory in everyday life , In U. Neisser (Ed) ,Memory observed : Remembering in natural contexts. San Fransisco : W.H.Freeman, 1982.

[3] B. Krudlwich: Lifestyle Finder. Intelligent User profiling using large-scale demographic data, AI Magazine, 18 (2), 1997.

[4] H. Lieberman: Integrating user interface agents with conventional applications, ACM Conf. on Intelligent User Interfaces, San Fransisco, USA, Jan. 1998

[5] H. Lieberman: Beyond information retrieval: information agents at the MIT Media Lab, Kunstliche Intelligenz, pp. 17-23, Mar. 1998.

[6] D. Fensel, Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, Second edition • • • • Berlin• Springer , 2004.

[7] M.aciej Ceglowski. Www::blog::identify - identify blogging tools based on url and content. http://search.cpan.org/ mceglows/WWW-Blog-Identify-0.06/

[8] T. Yamashita, and Y. Matsumoto: Language Independent Morphological Analysis, ANLP2000, pp. 232–238, 2000.

[9] J. Kleinberg. Bursty and hierarchical structure in streams. In Proc. of the 8th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp. 1–25, 2002.

[10] M. Okumura, and H. Mochizuki: Query-Biased Summarization Based on Lexical Chaining, Computational Intelligence 16 (4), pp.578–585, Nov. 2000.

[11] T. Kuraashima, T.Tezuka, and K.Tanaka: Proposal of Methods that Extract Topics in Town from Weblog, IEICE, 16th DataEngineeringWorkShop, Feb. 2005.

[12] T.Tezuka, R. yong Lee, H.Takakura, and Y. Kambayashi: The Image of the City Using Natural Language Analysis on Web Resources, DBSJ Letters Vol.1 No.1, Oct. 2002 (in Japanese).

[13] The special issue of "Easy to Use Practical Freeware for Natural Language Processing", IPSJ Magazine Vol.41, No.11, Nov. 2000.

IEEE COMPUTER SOCIETY