Information Gathering of Web pages to Guide Concept Understanding

YAMADA Seiji
CISS, IGSSE, Tokyo Institute of Technology
4259 Nagatsuta-cho, Midori
Yokohama 226-8502, Japan
+81-45(924)5217
yamada@ymd.dis.titech.ac.jp

OHSAWA Yukio
TOREST, Japan Science and
Technology Corporation
GSSM, Univ. Tsukuba, Tokyo, 112-0012 Japan
+81-3(3931)3718
osawa@gssm.otsuka.tsukuba.ac.jp

ABSTRACT

This paper describes navigation planning, a novel information gathering method that generates a plan for guiding concept understanding in the WWW. It also has the ability to generate operators during planning from Web pages. For understanding a concept, it is a useful way to utilize a search engine for gathering relevant Web pages. However the gathered Web pages are fractions of knowledge explaining a target concept directly. To cope with this problem, we propose navigation planning to generate a sequence of Web pages by which a user systematically understand a concept.

Keywords

Information gathering, AI planning

1. INTRODUCTION

Accessible information through the Internet increases explosively as the WWW becomes widespread. In this state, the WWW is getting useful for a user who wants to understand a *target concept* because he/she can browse helpful Web pages to understand the target concept. However, in general, this task is very hard because a user may not know where such Web pages are, and has to search them over the vast WWW search space. A practical solution to the problem is to utilize a search engine with the target concept as a query. However, since the retrieved Web pages are not filtered sufficiently, a user has to select useful ones from them. Furthermore, since in most cases the retrieved Web pages include concepts that a user does not understand, he/she must search the useful Web pages for them using a search engine again. Thus we consider Web pages obtained by a search engine are fractions of knowledge explaining a target concept directly. This browsing task is repeated until a user understands the target concept, and wastes time. We deal with the task as planning, and propose *navigation planning*[4] to automatically generate a sequence of Web pages which can guide a user to systematically understand a target concept.

2. NAVIGATION PLANNING

navigation means a task that indicates useful Web pages to a user for guiding his/her concept understanding. The sequence of Web pages is called a *plan*, and *navigation planning* means automatic generation of the plan. We can summarize the task in the following. This procedure is iterated until terminated by the user.

1. Search Web pages using a search engine.
2. Understand the pages retrieved by the search engine.
3. Select unknown concepts in the Web pages.
4. Go to *Step*1 with unknown concepts as target concepts.

We consider the above procedure classical planning[1] by the following correspondence.

- *Action*: Understanding concepts in a Web page.
- *State*: A user's knowledge state described with a set of words describing concepts which he/she knows.
- *Initial state*: A user's initial knowledge state.
- *Goal state*: A target concept described with a set of words which a user wants to understand.
- *Operator*: *U-Op(URL)* defined by the followings.
  - *Label*: URL of the Web page
  - *Condition*: $C = \{c_1, \cdots, c_i\}$, where C means the *condition words* which are necessary to understand the pages.
  - *Effect*: $E = \{e_1, \cdots, e_j\}$, where E is *effect words* which a user obtains by understanding the page.

This framework contains a significant problem which has not been dealt in planning. It is that the *U-Op(URL)* operators are not given in advance. Because it is impossible for a human designer to generate the operators from all Web pages in the WWW. Hence the operators need to be automatically generated from Web pages when they are necessary.

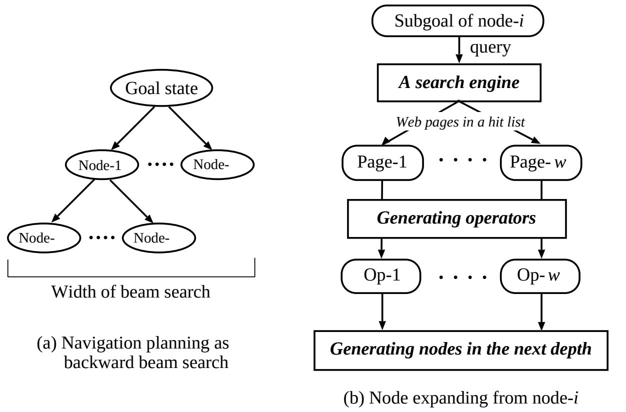
2.1 Automatic operator generation

We develop a method to extract condition and effect words from a Web page for generating an operator. Since the words are assumed to be written in a Web page, the problem is how to extract them from a Web page.

Using TAG structure

Various methods to extract keywords from text have been studied[3]. Though most methods are based on the occurrence frequency of words, one of the most effective methods is to utilize the structure in text. Since a Web page is described in a HTML format , we can utilize TAG structures.

Prime candidates for condition words are words linked to other Web pages, i.e. the words between `<A HREF=URL>`



(a) Navigation planning as backward beam search

(b) Node expanding from node-*i*

Figure 1: Navigation planning.

and `</A>`, because this tag is a sign of reference to relevant topics, which are important for understanding the current Web page in many cases.

Since a title in a Web page describes words which a user can acquire by reading the page, the words between `<TITLE>` and `</TITLE>` are candidates for effect words. In the same way, headings describe knowledge which a user can obtain by reading a section. Thus the words between `<Hn>` and `</Hn>` are also candidates for effect words.

KeyGraph: A keyword extraction method

Extraction of condition and effect words using only tag is not sufficient. Because not all linked words are condition words, and not all condition words are linked. Thus we need another method to assist it, and *KeyGraph* is selected.

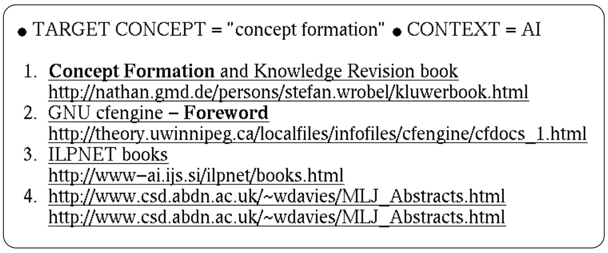
KeyGraph is a fast method for extracting keywords representing the asserted core idea in a document[2]. *KeyGraph* composes clusters of terms, based on co-occurrences between terms in a document. Each cluster represents a concept on which the document is based (i.e. condition words), and terms connecting clusters tightly are obtained as author's assertion (i.e. effect words). Furthermore the likelihood for condition and effect words can be computed by *KeyGraph*, and used for weight of an operator. Another merit of *KeyGraph* is that it does not employ a corpus.

The extraction of condition and effect words using tag structure and *KeyGraph* are integrated.

2.2 Planning procedure

We develop navigation planning procedure. Fig.1 shows the overview of navigation planning. It uses *backward beam search* from a goal state (Fig.1(a)). The node expansion (Fig.1(b)) includes the search for related Web pages with a search engine and the generation of operators.

We fully implemented a navigation planning system. Fig.2 shows a plan (depth = 4) with the target concept "concept formation". In this plan, the first page directly explains the target concept, and the third and fourth page explain "ILP" and "MLJ" which are unknown concepts included in the first page.



● TARGET CONCEPT = "concept formation" ● CONTEXT = AI

1. **Concept Formation** and Knowledge Revision book
http://nathan.gmd.de/persons/stefan.wrobel/kluwerbook.html
2. GNU cfengine **– Foreword**
http://theory.uwinnipeg.ca/localfiles/infofiles/cfengine/cfdocs_1.html
3. ILPNET books
http://www−ai.ijs.si/ilpnet/books.html
4. http://www.csd.abdn.ac.uk/~wdavies/MLJ_Abstracts.html
http://www.csd.abdn.ac.uk/~wdavies/MLJ_Abstracts.html

Figure 2: Plan for "Concept formation".

3. EXPERIMENTAL EVALUATIONS

We made experiments for evaluating our navigation planning. Using human subjects and their selected target concepts, we investigated the number of Web pages which were necessary for a user to understand a target concept. As a result, we verified navigation planning outperformed a popular search engine MetaCrawler.

Furthermore we made experiments for investigating advantage of planning over gathering fractions of knowledge like Web pages using a search engine. In these experiments, more difficult target concepts were selected to emphasize the difference of them, and gradual increase of user's knowledge with reading each page was investigated. As a result, we found out navigation planning is significantly more effective than a search engine in the view of systematic and deep understanding of a concept.

By these experimental results, we conclude that navigation planning contributes to increase user's knowledge until he/she deeply understands a goal concept, by reading multiple but small number of pages.

4. CONCLUSION

We proposed navigation planning, a information gathering method that generates a plan guiding a user to understand a concept in the WWW. It also has ability to generate operators during planning from Web pages. Search for useful Web pages for a user to understand goal concepts was formalized in a planning framework, and an operator corresponding to understanding of a Web page was defined with condition and effect words. Then we described the whole planning procedure and verified effectiveness of navigation planning experimentally.

5. REFERENCES

[1] R. E. Fikes and N. J. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2:189–208, 1971.

[2] Y. Ohsawa, N. E. Benson, and M. Yachida. *KeyGraph*: Automatic indexing by co-occurrence graph based on building construction metaphor. *IEEE Advanced Digital Library Conference*, pages 12–18, 1998.

[3] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Readings in Informationc Retrieval*, pages 323–328. Morgan Kaufmann, 1997.

[4] S. Yamada and Y. Osawa. Planning to guide concept understanding in the WWW. *AAAI 1998 Workshop on AI and Information Integration*, pages 121–126, 1998.