Navigation Planning to Guide Concept Understanding in the World Wide Web

Seiji Yamada
CISS, IGSSE
Tokyo Institute of Technology
4259 Nagatsuta, Midori
Yokohama 226-8502, Japan
yamada@ymd.dis.titech.ac.jp

Yukio Ohsawa
Department of Systems and Human Science
Tsukuba University
3-29-1 Ohtsuka, Bunkyou
Tokyo 112-0012, Japan
osawa@gssm.otsuka.tsukuba.ac.jp

ABSTRACT

This paper describes navigation planning 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 browse for relevant Web pages in the WWW. However this task is very hard because the user has to search for them in the vast WWW. To deal with this problem, we propose navigation planning to generate a sequence of Web pages by which a user systematically understand a concept.

1. INTRODUCTION

The WWW is very useful for a user who wants to understand a *target concept*. He/she can browse helpful Web pages to understand a target concept. However, in general, this task is very hard because he/she may not know where such Web pages are, and has to search them over the vast WWW search space. A solution of the problem is to use 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. This task is repeated until a user understands the target concept, and wastes time. We consider the task as planning, and propose *navigation planning* [4] to automatically generate a sequence of Web pages which can guide a user to understand a target concept.

2. NAVIGATION PLANNING

In this research, *navigation* means a task that indicates useful Web pages to a user for guiding his/her concept understanding. A sequence of useful Web pages is called a *plan*, and *navigation planning* means the 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.

The procedure above is considered *planning* [1] using the following correspondence. Using this formalization, we can apply a classical planning framework to navigation planning.

- *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.

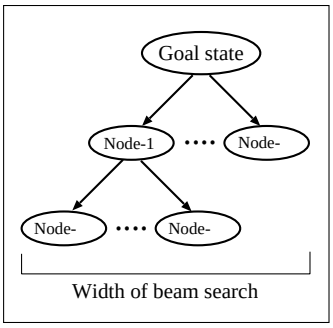
This navigation planning contains a significant problem which has not been in planning thus far. It is that the *U-Op(URL)* operators are not given in advance. This is because it is impossible for a human designer to generate the operators from all the Web pages in the WWW. Hence the operators need to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

3. GENERATING OPERATORS

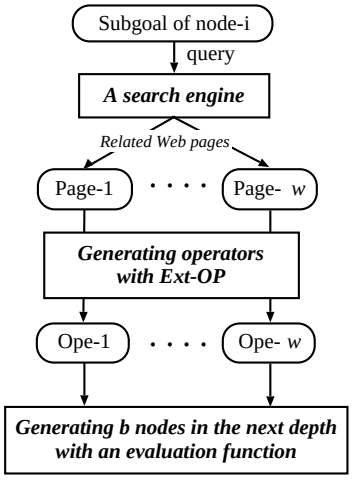
3.1 Using TAG structure in a html file

Various methods to extract keywords from text have been studied[3]. Though most methods are based on the 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.

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



(a) Navigation planning as backward beam search



(b) Node expanding from node-i

Figure 1: Navigation planning.

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

3.2 KeyGraph: A keyword extraction method

The extraction of condition and effect words using only the tag structure is not sufficient. All the linked words are not candidates for condition words, and all the condition words are not linked. Thus we need to utilize another method to assist it, and *KeyGraph* is used.

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.

4. PLANNING PROCEDURE

We now develop navigation planning procedure. Fig.1 shows the overview of our autonomous navigation planning agent, called *NaviPlan*. *NaviPlan* autonomously generates a plan for given target concepts. 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 *NaviPlan* and made various experiments with subjects for evaluation. Unfortunately, due to space constrains, we omit them. Fig.2 shows a plan (depth = 4) generated by *NaviPlan* with the target concept "concept formation".

5. CONCLUSION

We proposed a novel navigation planning method that generates a plan guiding user to understand a concept in the WWW. It also has the ability to generate operators during planning from Web pages using keyword extraction methods. The search for useful Web pages for a user to understand goal concepts was formalized using a planning framework, and an operator corresponding to the understanding of a Web page was defined with condition and effect knowledge. Then we described the whole planning procedure.

6. 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. In *IEEE Advanced Digital Library Conference*, 12–18, 1998.

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

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

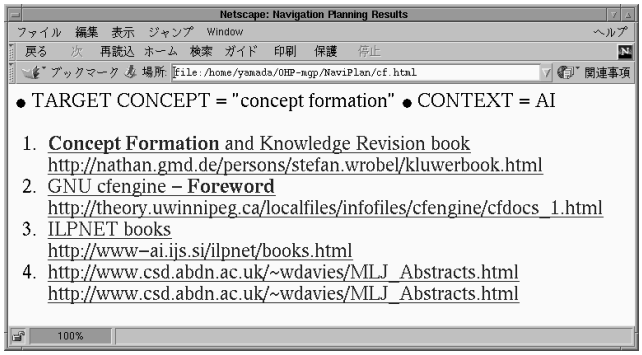


Figure 2: Plan for "Concept formation".