WebPlan: Dynamic Planning for Domain-Specific Search in the Internet

Jochem Hüllen & Ralph Bergmann & Frank Weberskirch
University of Kaiserslautern, Dept. of Computer Science
P.O. Box 3049, D-67653 Kaiserslautern, Germany
E-mail: {huellen|bergmann|weberski}@informatik.uni-kl.de

Abstract

Searching in the Internet is a dynamic, iterative process which has to be planned. Current available search engines are not able to do this satisfactory. On the other side classic action planning systems are not applicable to this problem directly. In this paper we present the goal of the *WebPlan* project which is dynamic planning for domain-specific search in the Internet. For this we extend the planning system *CAPlan* to build domain-specific search agents. ...

1 Introduction

During recent years, the World Wide Web (WWW) emerged to one of the most important sources of up-to-date information about nearly every topic, as well as a huge repository of software, multi-media, and other resources. Although almost anything is available somewhere, one of the biggest problems today is to locate the right resources or piece of information one is interested in. With the expected further growth of the Web this problem will get dominant or even might prevent the acceptance of Web technology by the larger group of non-expert users. Although current available search engines on the Web are of some use, their results are often not satisfactory. The main reason for this is that search engines primarily apply syntactic textual retrieval mechanisms. Experienced Web users know how to apply such search engines appropriately during a search process and, in addition, also apply lots of general domain knowledge and search strategies to achieve successful search results. Analyzing search strategies of expert Web users discloses that Web search often becomes a complicated process in which classical search engines are used as a single step executed as part of a more sophisticated search plan. The basic problems that make Web search hard can be summarized as follows:

- The information sources in the Internet are distributed, are extremely heterogeneous, and have different structures.
- Retrieving information from the Internet has a dynamic nature, because servers can be temporarily inaccessible, sources can be moved to other locations or could get even removed completely.
- The users queries are often incomplete and need to be reformulated several times.
- Although the Web resources are machine readable they are often not understandable for a machine.

- Some amount of domain knowledge is required to interpret the retrieved information and to decide whether it is appropriate or not.

Researchers from different areas have started to adress these problems by constructing information gathering systems that automatically query multiple, relevant information sources (Etzioni and Weld, 1994; Arens et al., 1996b; Levy et al., 1996; Kwok and Weld, 1996; Golden et al., 1996; Friedman and Weld, 1997; Singh, 1998).

The *WebPlan* project [1] (funded by the German Science Foundation, DFG) which started in August 1998 aims at developing techniques for building domain-specific search assistants for the Internet. The goal is to support the whole search process by looking at it as a goal directed, dynamic planning and plan execution problem. Since developing a universal domain-independent retrieval assistant does not seem feasible because it would have to deal with a huge amount of common sense knowledge, we focus on domain-specific search. This enables considering some domain specific knowledge in order to recognize the semantic and pragmatic aspects of a user query to some degree. In the project we will primarily focus on the domain of PC software, i.e., finding an appropriate piece of software from the Web (e.g. a Windows95 driver for a particular graphics card).

This paper describes the preliminary architecture of a domain specific search assistant that will be refined and implemented in the course of the project. It shows how this work will extend the *CAPlan* architecture (Weberskirch, 1995; Weberskirch and Paulokat, 1995) developed at the University of Kaiserslautern in recent years.

We will illustrate the planning aspect of systematic domain-specific internet search, first by giving a detailed example and second by introducing the process by a general model. Then, section 4 gives some details about query formulation and interpretation before section 5 describes all important aspects of planning for information gathering. In section 6 the *WebPlan* architecture will be presented.

2 Example for Domain-Specific Planned Search

The following example shows a typical systematic search process in the domain of PC software. It illustrates how a user who is familiar with searching the Internet looks for a driver for a graphics board. It demonstrates that this search process is in fact a goal directed planning process.

Suppose that an experienced Internet user is searching for a driver for his new graphics board *miroVideo22SD* for *WINDOWS 95*. On an abstract level he usually has the following plan:

1. *Find an URL which offers a driver for the* miroVideo22SD *graphics board.*
2. *Copy this driver to the local harddisk.*

There are several possibilities to refine the first step. One possibility is to find an HTML page that contains the strings: *miro*, *Video* and *22SD*. This plan may not have a high probability of success, but it could be worth trying because it is very simple. A possible refinement is this action sequence:

1.A.1 Choose a conventional search engine.
1.A.2 Construct a query for this machine.

[1] http://wwwagr.informatik.uni-kl.de/~webplan

1.A.3 Follow the retrieved links and check if they offer the required driver.

The execution of these actions lead to the following results:

1.A.1 The search engine AltaVista *was choosen.*
1.A.2 The query miro AND Video AND 22SD *is entered in the advanced mode.*
1.A.3 No links have been found.

This execution doesn't solve the problem but it gives some new information: Since no suitable links were found it would be useful to pose a generalized query. Hence, the following steps are added to the plan:

1.A.4 Generalize the query.
1.A.5 Follow the retrieved links and check if they offer the required driver.

The execution of the added actions had the following effects:

1.A.4 The query miro AND Video *has been asked in the advanced mode.*
1.A.5 More than 8000 links were returned by the search engine. The first hits point to pages containing videos about the surrealist Jean Miro [2].

The 8000 links lead to a huge amount of information and there were even a few links which lead to information about graphics boards from *miro* but it was not possible to find a link to a suitable driver. Because this first attempt failed, a second more complicate plan is developed:

1.B.1 Find out the manufacturer of the graphics board.
1.B.2 Search for information sources from the manufacturer.
1.B.3 Localize the web area about graphics driver.
1.B.4 Localize the driver for the miroVideo22SD *board.*

The execution of theses actions lead to the following:

1.B.1 miro *is the manufacturer of the graphics board.*
1.B.2 http://www.miro.com *is the manufacturer's homepage.*
1.B.3 On this page there is a link to graphics driver.
1.B.4 There is a link to a file 243711.exe *which is the driver for the* miroVideo22SD *board for* Windows 95.
2. Copy 243711.exe *to the local harddisk.*

This example illustrates the following general planning issues:

- Domain-specific internet search is a dynamic process. New information is determined during this process which influences further (planning) activities.
- This process is goal directed; in the example the user is not interested in general information about graphics boards or drivers - he wants one specific driver.
- Domain specific background knowledge is required (like in step 1.B) in different ways.
- AI Planning is an appropiate approach to support this process provided its is able to deal with the dynamics and integrates smoothly with plan execution and monitoring.

[2] This is a typical result for search engines because they only consider syntactic matches but ignore the semantic or pragmatic aspects

3 Process of Systematic Domain-Specific Search

Generalizing the example presented in the previous section, we can identify the cycle for systematic domain-specific internet search shown in Fig.1. This cycle is a modification of the knowledge-based information seeking process in (Carranza and Lenski, 1997).

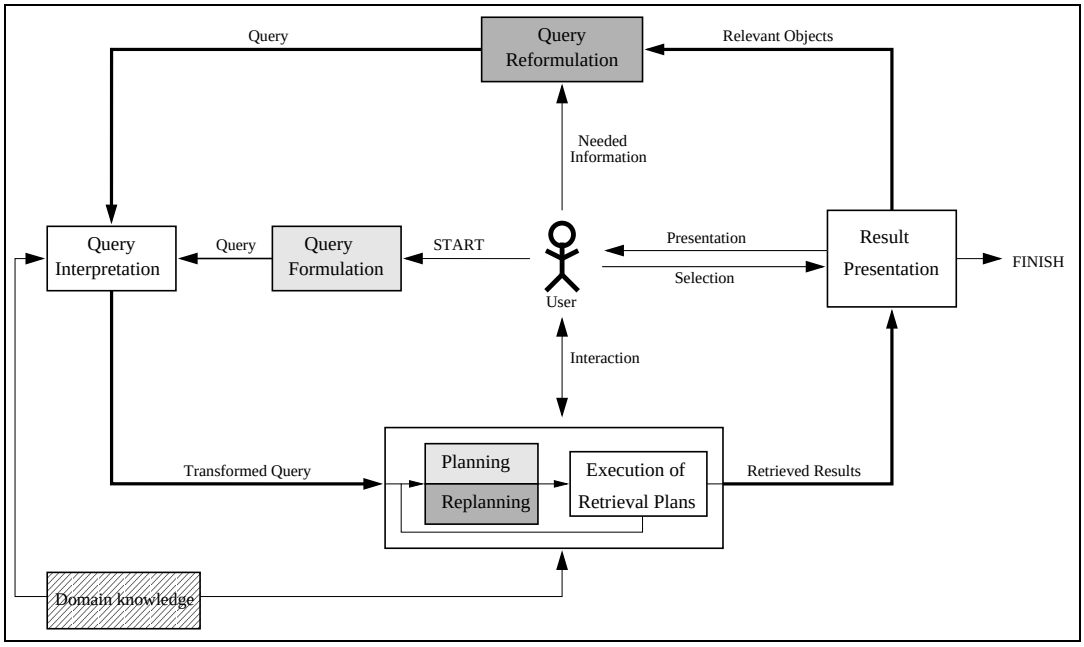

Figure 1: Planning-based domain-specific internet search

In this cycle the user plays an important role in most of the following phases (the importance of the user is also described in (Belkin, 1996)):

- The user poses the initial domain-specific query to the system using a predefined form (**Query Formulation**). It should contain a description of the required information or resource. Often this initial query is vaguely stated, uncomplete or ill-defined. The query is expressed in the language offered by the system. This language should support the formulation of the semantic and pragmatic aspects of a query. Further details are described in the next section.

- In the **Query Interpretation** phase the query is analyzed and transformed. The domain knowledge is used to extract the semantic and pragmatic aspects and to transform the query into an internal representation based on a domain ontology. One part of the transformed query is the identified planning problem.

- **Planning and execution** of retrieval plans has to be interleaved because the planning of further steps could depend on the results of the execution of a retrieval operator. In this phase the domain knowledge is also needed, for example to select certain information sources. There could be interaction between the system and the user (e.g., the structure of some information sources is not completely known by the system, so the user has to help to interpret some retrieved results). Moreover the user should be able to monitor the whole planning process and especially the execution of the retrieval operators.

- In many cases the first **Result Presentation** will not satisfy the user, therefore, further runs through the cycle are necessary. So the user can select some objects and reformulate the query using the new knowledge. The search process will then start with the reformulated query.

- In the most cases the reformulated query is a modification of the first query but not totally different from it. So, the planning system should be able to replan starting with the first retrieval plan. In many cases it would be unappropiate to plan from scratch. *CAPlan*, where *WebPlan* will be built on, supports **Replanning** since it uses the dependence maintenance system *REDUX* (Petrie, 1992).

4 Query Formulation and Interpretation

The query for classical search engines is normally a list of some keywords. Most search engines support several boolean functions to connect these keywords. Why is the response to such queries often not satisfactory (typically there are no matches or hundreds or thousands of matches and the most interesting are possibly missing)? The main reason for this is that only syntactic matches are performed, but there are no possibilities to consider the semantic and pragmatic aspects of the users query. It would be rather difficult to embed such aspects in universal search engines, because the system would have to deal with a lot of common sense knowledge. In domain specific search, however, it is possible to overcome this problem since domain-specific background knowledge could lead to the real purpose of the query. Furthermore, in a domain specific scenario, the possible semantics and pragmatics a user might have can be analyzed in advance.

Specific query forms can be developed that capture these predefined aspects. Moreoever, it is usefull that the system guides the user during the formulation of the query. For the WebPlan system we want to use domain-specific hierarchically structured query schemes. The user can fill some of the fields of these schemes with free text and others by choosing keywords from defined lists. So, the system is able to recognize the semantic and pragmatic aspects of the query. Due to the cycle mentioned above the module for the query input should support mechanisms to reformulate queries using objects from the results of the previous queries.

The purpose of the Query Interpretation phase is to construct a planning problem. In general the planning problem is to achieve a goal state in which needed information is retrieved. Executing this plan performs the actual retrieval operations. Therefore, the purpose of the query has to be detected and the query has to be transformed into a set of information goals. For this purpose, a lot of domain knowledge (the domain ontology and the knowledge about the information sources) has to be considered.

Beside the domain-specific operators a set of domain independed information goals and information-seeking actions have to be defined. The analyzed information goals (domain dependent or independent) have to ensure that the execution of a solution of the planning problem consisting of these goals answers the users query in a satisfactory manner.

5 Planning for Domain-Specific Internet Search

Since information gathering is a dynamic goal-directed process and domain-specific search needs domain-specific background knowledge, action planning is an appropriate method. Though a classical AI planning system has to be extended with several features. WebPlan will be built on top of CAPlan, a SNLP-based (McAllester and Rosenblitt, 1991) plan-space planner. For this approach the following aspects have to be considered:

Incomplete Knowledge: Classical AI planners assume complete knowledge about the start situation of a planning problem. When dealing with information gathering not only the needed information is unknown. Also information like the status of a particular server or which objects an information source contains is unknown. So the system has to deal with incomplete knowlede about the state of the world. One consequence is that the CWA (closed world assumption) does not hold. Hence, if a certain fact cannot be deduced by the system this does not mean that this fact is not true. Instead, the system has to decide whether it should try to get information about it. To support such decisions it is useful to annotate the preconditions and effects of the operators to characterize them: The preconditions can be divided into those that should be deduced and those that should be satisfied by seeking for new information. Similarly, the effects can be devided into those that change the world and those that change the system's knowledge about the world. These methods have been discussed in (Etzioni et al., 1992; Golden et al., 1996). To overcome the problem of missing CWA one can represent facts that are *false*, explicitly. Another approach is to use LCWA (local closed world assumption (Golden and Weld, 1996)). This means that the system uses LCW in single fields (e.g. an information seeking operator surely gives all objects with certain attributes, so the system knows that all other objects do not have these attributes).

Information Seeking Operators: To retrive new information some information seeking operators a necessary. To integrate them in the system it is usefull to merge the domain-specific operators with some general information seeking operators. To refer the retrieved information later on, one possibility is to use run-time variables (described in (Golden and Weld, 1996; Knoblock, 1996)). They are bound when an information seeking operator is executed and contain the retrieved information.

Planning and Execution: Usually, a solution to a planning problem is a *plan* which transforms the start situation into the finish situation. The execution of such a plan is a task for another system (e.g., a robot). In contrast to this a solution of an information gathering planning problem is the *execution* of a plan which solves the problem. Since there are operators which retrieve new information, it is possible that other decisions of the planning system depend on the result of those information seeking operators (in the introducing example the execution of the operator 1.A.3 causes the system to use operator 1.A.4). So it is necessary to interleave planning and execution of information seeking operators. Consequently, one needs a combined system for planning and execution. Interleaving planning and execution causes two problems: First the system has to decide which operator should be executed at which time (several operators could be executed simultaneously and planning and execution could be done parallel). Secondly, it is not possible to backtrack over executed operators. So the decision to execute an operator has to be made

very carefully. If only operators which change the knowledge about the world but which don't change the world itself are executed, there would be no problems with effects which are impossible to reverse but there could arise unnecessary costs.

Plan-Space/State-Space: A consequence of interleaving execution and planning is the necessity to modify the general planning algorithm. Since the worlds state could be changed by some executed operators during planning, the plan-space based planning algorithm should be combined with a state-space approach, like in UCP (Kambhampati and Srivastava, 1995; Kambhampati et al., 1998). An important difference to UCP is, as mentioned above, that it is not possible to backtrack over the state-space plan refinements because they are already executed.

User Interaction: As mentioned before, the user plays an important role in the process of information gathering. There are several points in the cycle where the user interacts with the system. So, the planning system has to be extended to support such interaction. The user should be able to control the planning process. Hence the control component of the planning system has to present the current state of planning to the user. Of course the system should be able to plan independently but the control component should also allow the user for example to prefer some operator in a certain situation or to start the execution of an operator. Especially, if there is not enough knowledge about some information source (in the example for the step 1.B.3) the user may help to interpret some results (this could also be done by a text identification system but this is not the goal of the *Webplan* project).

Hierarchical Planning: In the introducing example there are abstract and concrete operators (1. and 1.A.1). This is typical for information gathering domains. So, a hierarchical planning approach seems to be usefull. *WebPlan* will use an extended version of *CAPlan* that supports hierarchical task decompositions (similar to *DPOCL* (Young et al., 1994) and *UCP* (Kambhampati et al., 1998)) combined with *SNLP* and dependence maintenance for interactive planning.

Information Sources: The advantage to use the Internet as an information source is to have very much information about nearly all domains. On the other hand the several sources in the Internet have no standardized structure. So it is usefull to know the location and the structure (format of query and answer) of some sources which are relevant for a specific domain. In a further state the *WebPlan* system should also be able to determine relevant sources. Of course, the system doesn't know the structure of these sources exactly. One approach is to involve the user in the problem of interpreting the retrieved information. Another possibility in the future will be to do this automatically since there are a lot of aspirations to comment web documents in a standardized form (Luke et al., 1997).

6 Architecture of the *WebPlan* System

The *WebPlan* system will be built on top of *CAPlan*. So we extend this system for classic action planning with the features mentioned in the previous sections. Figure 2 illustrates the architecture of the *WebPlan* system, which is connected to the Internet via a WWW-Server. Most components are generic, they realize the basic mechanism for planning-based

internet search. The grey components are domain-specific, they contain the knowledge and the code for actions in the concrete application.

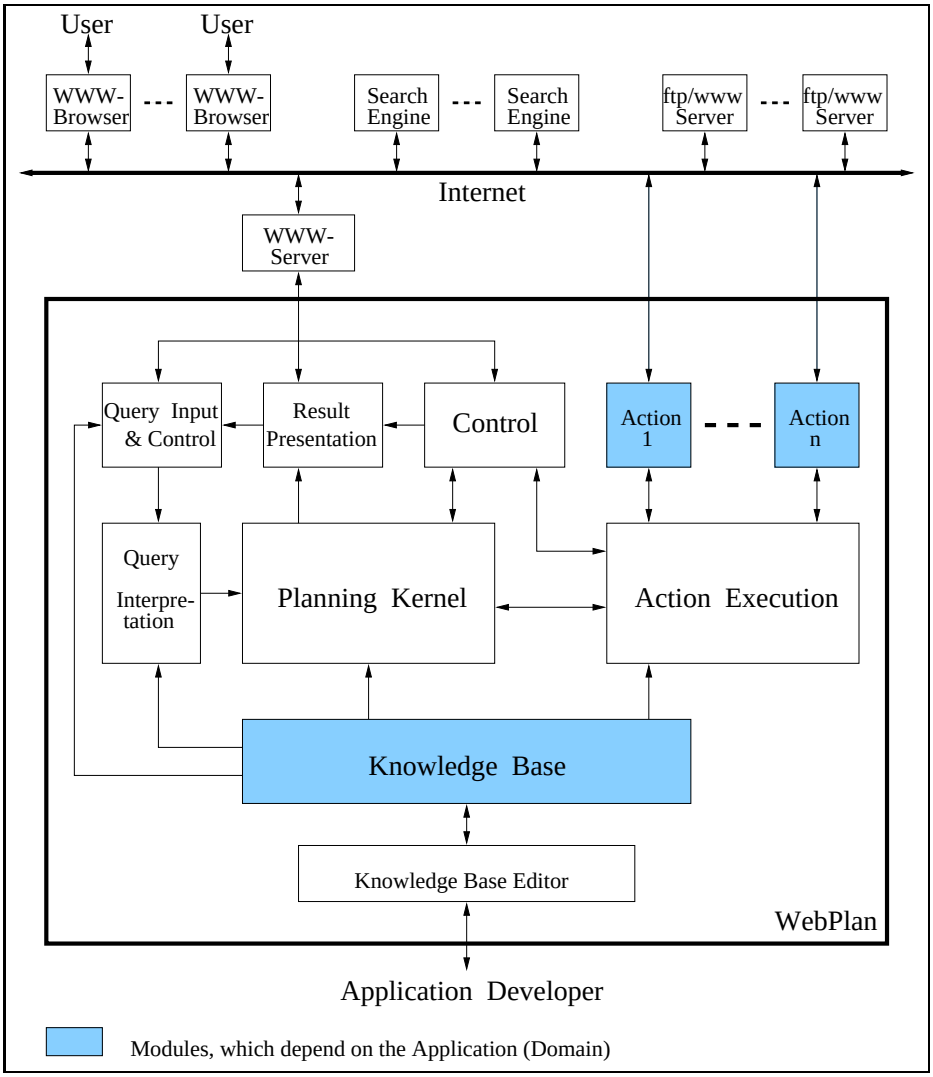

Figure 2: The Architecture of the WebPlan system

Planning Kernel: This module realize the basic planning algorithm for domain-specific internet search. It can be seen as an extension of the planning kernel of *CAPlan*. These extensions were discussed in the previous section.

Action Execution: This component is responsible for controlling the execution of actions. It interacts frequently with the planning kernel which determines operators that should be executed. The execution module monitors the execution of the actions which correspond to the operators and gives the results to the planning kernel.

Query Interpretation: In this module the users query has to be tranformed into a planning problem, this was discussed in section 4.

Knowledge Base: The knowledge base consists of three main parts:

1. The definition of the deductive and information seeking operators. This declarative knowledge could be compared with the domain definition in a classic planning system and is needed by the planning kernel.
2. The code for the domain-specific information seeking actions. This procedural knowledge is needed by the execution module.
3. The vocabulary and the mechanisms for the queries. This knowledge is mainly used by the Query Input, Control and Interpretation component.

Knowledge Base Editor: This editor should support the application developer to create and service the domains. It offers mecanisms to edit the three parts of the knowledge base mentioned above.

Control: This module controls the search process. It shouldn't be mixed up with the control component of the planning kernel which has to control the planning process. The control module of the *WebPlan* system has to monitor the search process and has to present the current state (current plan and results of executed actions) to the user, which is done via the Result Presentation module. The user should be able to interfere in the search process as described in the preceding sections.

Result Presentation: This component works up the search results and present them to the user. Retrieved information could be used to reformulate the initial query.

Query Input & Control: This module realize a WWW-User-Interface for the input and refinement of queries as well as for interactive control of the search process.

In the *WebPlan* project this general architecture will be refined and implemented. Further a domain-specific knowledge base for the domain of PC software will be built to realize a search agent for PC software in the Internet.

7 Conclusion

We presented an overview of the goal of the *WebPlan* project, which is the development of a generative, interactive tool for planning-based domain-specific search in the Internet. We'll construct a search agent for the domain of PC software and believe that in future the importance of such search agents will increase as the available information in the internet increases.

Another possibility for domain-specific search in the Internet is to build specialized applications by hand. These applications are time-consuming and costly to build, and difficult to maintain. The *WebPlan* system is a tool for creating rapidly domain-specific search agents.

There are other approaches for Information agents: *Internet Softbot* (Etzioni and Weld, 1994), *XII* (Golden et al., 1996), *SIMS* (Knoblock, 1996; Arens et al., 1996a), *Occam* (Kwok and Weld, 1996), *Razor* (Friedman and Weld, 1997), *Tessarae* (Singh, 1998), *Ariadne* (Knoblock et al., 1998)

The most closely related project is the *SIMS* project from *ISI*, in which also an AI planner is applied to the information seeking task. It focuses on the integration of well-structured databases. The *Ariadne* project from *ISI* deals with accessing information from more loosly structured Web sources.

In the *WebPlan* project first we are concerning with access from well- and semi-structured information sources. Later on we'll work at the problem that the system itself should determine information sources and so there will be very few knowledge about such sources. Further we expect advantages from two aspects of the planning system *CAPlan*:

1. The hierarchical planning approach offers comfortable possibilities to define the domain-dependent and -independent aspects of the search agents. The use of abstract operators is natural in such domains.

2. The use of the dependence maintenance system *REDUX* supports replanning which is important when dealing with query planning (as mentioned in section 3).

Future work could focus on applying case-based planning for Internet search. This would lead to an extension of the case-based component *CAPlan/CbC* (Muñoz-Avila, 1998) of *CAPlan*.

References

Arens, Y., Hsu, C.-N., and Knoblock, C. A. (1996a). Query processing in the sims information mediator. In Tate, A., editor, *Advanced Planning Technology*. AAAI Press.

Arens, Y., Knoblock, C., and Shen, W. (1996b). Query reformulation for dynamic information integration. *Journal of Intelligent Information Systems*, 6(2/3):99–130.

Belkin, N. (1996). Intelligent information retrieval: Whose intelligence? In *Proceedings of 5th Intern. Symposium on Information Science (ISI-96)*.

Carranza, C. and Lenski, W. (1997). Planning the information seeking process in heterogeneous bibliography-related sources. In *Workshop KI-Techniken für intelligente Mensch-Maschine-Schnittstellen, 21. Deutsche Jahrestagung für KI (KI-97)*.

Etzioni, O., Hanks, S., Weld, D., Draper, D., Lesh, N., and Williamson, M. (1992). An approach to planning with incomplete information. In *Proceedings of KR-92*.

Etzioni, O. and Weld, D. (1994). A softbot-based interface to the internet. *Comm. of ACM*.

Friedman, M. and Weld, D. (1997). Efficiently executing information-gathering plans. In *Proceedings of IJCAI-97*.

Golden, K., Etzioni, O., and Weld, D. (1996). Planning with execution and incomplete information. Technical Report UW-CSE-96-01-09, Dep. of Computer Science and Engineering, University of Washington.

Golden, K. and Weld, D. (1996). Representing sensing actions: The middleground revisited. In *Proceedings of KR-96*.

Kambhampati, S., Mali, A., and Srivastava, B. (1998). Hybrid planning for partially hierarchical domains. In *Proceedings of AAAI-98*.

Kambhampati, S. and Srivastava, B. (1995). Universal classical planner: An algorithm for unifying state-space and plan-space planning. In *Proceedings of the 3rd European Workshop on Planning (EWSP-95)*.

Knoblock, C. (1996). Building a planner for information gathering: A report from the trenches. In *Proceedings of the 3rd Intern. Conf. on AI Planning Systems (AIPS-96)*, pages 134–141.

Knoblock, C. A., Minton, S., Ambite, J. L., Modi, N. A. P. J., Muslea, I., Philpot, A. G., and Tejada, S. (1998). Modeling web sources for information integration. In *Proceedings of the 15th National Conf. on Artificial Intelligence*.

Kwok, C. and Weld, D. (1996). Planning to gather information. In *Proceedings of AAAI-96*, pages 32–39.

Levy, A., Rajaraman, A., and Ordille, J. (1996). Query-answering algorithms for information agents. In *Proceedings of AAAI-96*, pages 40–47.

Luke, S., Spector, L., Rager, D., and Hendler, J. (1997). Ontology-based web agents. In *Proceedings of the 1st International Conf. on Autonomous Agents.*

McAllester, D. and Rosenblitt, D. (1991). Systematic nonlinear planning. In *Proceedings of AAAI-91*, pages 634–639.

Muñoz-Avila, H. (1998). *Integrating Twofold Case Retrieval and Complete Decision Replay in CAPlan/CbC.* PhD thesis, University of Kaiserslautern.

Petrie, C. (1992). Constrained decision revision. In *Proceedings of AAAI-92*, pages 393–400.

Singh, N. (1998). Unifying heterogeneous information models. *CACM*, 41(5):37–44.

Weberskirch, F. (1995). Combining SNLP-like planning and dependency-maintenance. Technical Report LSA-95-10E, Centre for Learning Systems and Applications, University of Kaiserslautern, Germany.

Weberskirch, F. and Paulokat, J. (1995). CAPlan - ein SNLP-basierter Planungsassistent. In Biundo, S. and Tank, W., editors, *Beiträge zum 9. Workshop 'Planen und Konfigurieren' (PuK-95)*, number DKFI-D-95-01 in DFKI-Dokument. DFKI.

Young, R., Pollack, M., and Moore, J. (1994). Decomposition and causality in partial-order planning. In *Proceedings of the 2nd Intern. Conf. on AI Planning Systems (AIPS-94)*, pages 188–193.