An Advanced system

for Information retrieval via key concepts

Hiroyuki KAMEDA[1], Noriko OOMORI[2], Chiaki KUBOMURA[3] and Yukinobu TANIFUJI[4]

School of Information Technology,

Faculty of Engineering,

Tokyo University of Technology,

1404-1 Kataura, Hachioji, Tokyo 192-0982, JAPAN

{kameda[1], tanifuji[4]}@cc.teu.ac.jp, non@ma.it.teu.ac.jp[2], kubomura@ed.teu.ac.jp[3]

Abstract

Recently, many advanced information retrieval systems have been intensively studied for the WWW (World-Wide Web) on the Internet. The current information retrieval systems in use, however, have not yet enough abilities, so that users can swiftly retrieve just what they really want exactly with no data losses and no noises.

In this paper, we present an overview of a new advanced information retrieval system based on a information retrieval model. The advanced information retrieval system, which is not completely implemented on computers so far, have abilities such as "information retrieval via key concepts" based on an information structure for information retrieval, "unknown word acquisition," "intention understanding," "knowledge acquisition" for user modeling and data configuration etc. At first, we briefly describe an overview of the previous information retrieval systems and their problems. Then our new model of information retrieval is described to cope with the problems. Some cores of modules of the model are also described in some details. Especially, the method: "information retrieval via key concepts" is intensively described, which is based on information structure for information retrieval. As our information retrieval system was partially evaluated with use of the NACSIS data, the evaluation result is also reported briefly. In concrete, some problems of our system which was clarified through the evaluation with use of the NACSIS data are discussed, and new problems that our system mainly must cope with next are also presented as the results and conclusions of the system performance evaluation.

Key words

information retrieval via key concepts, information structure for information retrieval, knowledge acquisition, Evaluation with use of NACSIS data

1. Introduction

As computer technologies have made a remarkable progress, the construction of information systems in the form of network and the distributed management of information are rapidly prevailing in the society. Especially, much amount of information is affirmatively publicized to the Internet society via the Internet, due to constructing the Internet systematically and broadly in the world. But no matter how much amount of information we gather, store on the Internet, and publicize to the society via the Internet, it is no use constructing information retrieval systems on it, if they cannot swiftly provide appropriate pieces of information for people who need the information (hereafter "user").

From this point of view, we may say that wide range and much amount of information is now in Japan available for everyone, because the information is affirmatively publicized by many people, such as the government, academic researchers, housekeepers, even elementary school children. No solutions to the problem that the swift and appropriate information retrieval are, however, not yet given, in spite that some information retrieval systems, e.g. YAHOO! and Infoseek are available for the use now.

In this report, we present a new advanced information system on the Internet and its evaluation results with use of NACSIS data. In the section 2, the previous information retrieval systems and their problems are described. And in the sections 3 and 4, facilities for Advanced Information Retrieval Systems, and Overview of New Advanced Information Retrieval System are presented and discussed. We also describe and considered an evaluation of our new advanced information retrieval system in the section 5. Finally, conclusions are described in the last section 6.

2. The Previous Information Systems and their Problems [1]

2.1 The previous information retrieval systems

In order to make our new advanced information retrieval system understandable, the previous information retrieval system is described in the following. In this report, the previous information retrieval system means to be the processes:

-A user explains what he/she wants to get to a information retrieval expert (hereafter "searcher" or "agent.") Sometimes the searcher is the user itself.

-The searcher extracts information from the places where it is stored.

-The searcher presents the information appropriately to the user.

In this section, the following words are defined as follows:

-Information: data files, books, etc.

-Places where information is stored: data file rooms, libraries, etc.

Information must be organized and stored in advance, so that information retrieval can be done swiftly and appropriately. Therefore, the previous information retrieval systems are in general configured with two parts: information storage subsystem and information retrieval subsystem.

2.1.1 Information Storage Subsystem

Information storage subsystem stores information materials as information retrieval objects, and has the following 4 facilities:

(1) Information Collection and Selection

Much amount of data are continuously collected from a wide range of fields to make database abundant in data. Collected data should be selected by filtering them in examining if the contents of the data are proper to be stored for the database.

(2) Information Analysis

The selected data are checked from the following point of views:

- whether the contents of data are correct,

- whether the data are not multi-registered,
- whether the data are brand-new or obsolete,
- to which category the data should belongs, etc.

(3) Filing of Information Materials

When filing data, accessibility is usually important. Data are edited to be in the database-proper format, and is grouped into every category to be stored as a file onto the database.

(4) Indexing

As Searching full database for a data is time-consuming even if sets of data are well-organized, indexing is necessary. For example, key words for search are attached to every filing material.

2.1.2 Information Retrieval Subsystem

Information retrieval subsystem fills the role of information extraction from stored materials. For a series of information retrieval processes from user's query to information presentation for the user, the following facilities are needed:

(1) Statement and Confirmation of query

When a user want to search for a data, he/she, at first, should state his/her own request clearly in the form of query. Dialogue process is usually inevitable to confirm queries, when an agent searches for data in place of the user.

(2) Decision of Search Strategy

The contents of queries given by the user are inferred and analyzed from every aspects to make a search plan (hereafter "search strategy.") For examples, search strategies are to establish relationships between words in the queries and key technical words for search, and to confine search space to smaller and more relevant one for ease of searching.

(3) Conversion to search words

According to the search strategies said above, the contents of queries given by the user are converted into expressions for search (hereafter "search words.")

(4) Extraction of Materials

All relevant data or files are extracted from the database (data files.)

(5) Matching between query and Extracted Materials

The contents of the extracted files are tested to see if they match with the4 contents of queries given by the user. When they do not match with each other, new search strategy is planned to make a successful search based on it.

(6) Presentation of Extracted Materials to user

Extracted files are presented to the user, when they match with the queries accurately. Ways of presentation, e.g., order of presentation, are seriously important, when the number of the extracted files are very enormous.

2.2 The Previous Information retrieval System on the Internet

In the information retrieval systems on WWW on the Internet, the following two information retrieval methods are usually adopted, and they have respectively the following merits and demerits said below.

(1) Directory Layer Method

In this method, database organizer itself constructs indexing files by checking the contents of data materials. Therefore, indexing is usually correct. But because all information distributed in all over the world, it is infeasible to collect all data from all over the world with no missing and no biasing. Moreover, it also cannot be helped to fail to keep track of changing data in real-time. For example, URLs (Uniform Resource Locators) on the Internet are subject to change everyday.

(2) Key Word method

Recently in this method, data are constantly collected with use of "WWW robot" from data on WWW. Therefore, information retrieval systems adopting this method contain immensely amount of data in systems. Moreover, data are kept up-to-date by replacing obsolete data with brad-new one. On the other hand, in this method, search is done by way of string of

key word input by a user. This leads to collecting some incorrect data caused by the existence of homographs to input key words, and also to miss some data caused by the existence of heterographs to input key words.

3. Facilities for Advanced Information Retrieval Systems [1]

In order to resolve the problems said before in this report, we present a new information systems with advanced facilities: sentence comprehension, unknown word processing, intention inference, information retrieval by way of key concepts, and knowledge acquisition.

(1) Sentence Comprehension Facility

Queries given by a user in Japanese sentence are processed in dialogue way, as well as Information in text (Japanese) form is also processed in non-dialogue way, to extract key words (hereafter "KW") and key concepts (hereafter "KC") from linguistic expressions in the queries and text data. Personal information of the user is also extracted, e.g., from his/her information retrieval behavior patterns.

(2) Unknown Word Processing Facility

Unknown words, which means words unregistered in the system dictionary, may appear in the dialogue between users and the systems, or in data stored on the Internet. In the sentence comprehension processing said before, this unknown word processing facility is invoked to process text with unregistered words in the system dictionary (hereafter "unknown words") smoothly.

(3) Intention Inference Facility

When user fails to express what he/she really wants to retrieve (hereafter "intention") in his/her own words explicitly, the intention of user's information retrieval query is inferred by the system automatically from discourse information and so on.

(4) Information Retrieval via Key Concepts Facility

Concepts conveyed by key words are extracted to retrieve relevant information via the key concepts. This facility reduces noises and misses in information retrieval, which are caused by existence of homographs and heterographs.

(5) Knowledge Acquisition Facility

Many kinds of knowledge for advanced information retrieval are acquired by the system. Such as unknown words in queries, user's habits in terms of pragmatics, dialogue histories used for intention inference, meta-knoweldge of characteristics of databases on WWW (World Wide Web) on the Internet, knowledge of effective combinations among key words and key concepts and so on. The more completely the facility is realized, the more autonomously and intelligently the system runs.

4. Overview of New Advanced Information Retrieval System[1]

In the following, we describe an overview of our new ADvAnced inforMAtion retrieval System (hereafter "ADAMAS-1") in some details.

4.1 Information Retrieval Engine

In the ADAMAS-1, the part of information retrieval engine, in principle, gives a facility of detecting linking relationships between link data and database URL. It also consists of two components IFA (Interface Agent) and IRA (Information Retrieval Agent.) IFA is a software agent for man-machine interface. IRA is a one for searching with use of linking data.

The information retrieval engine also has 4 modules of understanding dialogue between users and the system, generating key concepts corresponding to key words, generating search formulae, selecting data, as well as 2 knowledge of personal information and URL linking data.

4.1.1 Processing in Information Retrieval Engine

At first, searcher gives query to the system in Japanese sentences through keyboard. The system passes

the sentences to the dialogue understanding module to produce key words and key concepts. Then in order to confirm the correctness of key words and key concepts, the system presents the key words and key concepts in natural language form (Japanese), and ask the user whether the key words and key concepts are OK or not. If no OK, the user and the system make interaction in Japanese repeatedly until the user is satisfied with the key words and key concepts presented by the system. If OK, the user commands the system to execute to search all over files for appropriate data.

Next, the final key word generating module accepts the key words and key concepts said above to produce the final key words for search, and they are passed to the search formula generating module to match them with lining data(, which contains URLs and the content text of the files pointed by URLs.)

The results of information search are presented in a appropriate manner to the user through reducing doubly-picked data, filtering the data, and even odering the data when the number of data is huge by the data selection module.

When no appropriate results are extracted, search is done again from the beginning, and for the purpose that no such failure happens any more, the system keeps track of and store the search history, dialogue, and sets of key words and key concepts as system knowledge of personal information.

4.1.2 Dialogue Sentence Comprehension module

The dialogue understanding module consists of 4 modules and knowledges corresponding to the modules, i.e., morpheme analysis module, syntactic analysis module, meaning analysis module, and intention inferring module.

4.1.3 Key Word Generating Module

The key word generating module is a one to produce sets of candidate key words for practical search based on the model described later in the subsection 4.3.

4.1.4 Retrieval Formula Generating Module

The search formula generating module produces a search formula consisting of the logical combination (logical AND, logical OR, Negation) of the candidate key words given by the key word generating module.

4.1.5 IR Information Selection Module

The IR information selection module checks the doubly-picked data, and match the key words and key concepts in the linking data with the contents of what user wants to get.

4.1.6 Personal Information Module

The personal information module is implemented to make the dialogue understanding module run swiftly. This module collects and stores background knowledge of users (research fields, hobbies, titles, born place, and etc.) as personal information data. Dialogue history is also stored as personal information data to support intention understanding in dialogue between the user and the system appropriately.

4.2 Storage Engine

Storage engine is a subsystem for storing linking data onto the system. Storage engine has ICA (Information Collection Agent) as superviser agent to collecting data from database distributed in all over the world. ICA consists of 3 modules, i.e., information collecting module, document understanding module, and IR-meta information selection module, as well as a set of linking data as knowledge module.

4.2.1 Processing in Information Storage Engine

By invoking the information collection module, text data distributed in the world are collected through the Internet. The document understanding module processes the text data to produce a set of URL, key

concepts, and key words, which are stored as linking data on the system, while doubly-picked data are eliminated and some data are up-dated at any time if necessary.

4.2.2 Information Collecting Module

The information collecting module consists of programs called "WWW robots," "Spider," and "wanderer" in terms of the Internet. The module collects URLs and text data (document) by way of HTML (HyperText Markup Language) tags in the files.

4.2.3 Document Understanding module

The document understanding module accepts the text data from the information collecting module to produce the key concepts and key words of the text data, and store them with their URLs as system knowledge.

4.2.4 IR Meta-Information Selection Module

The IR meta-information selection module checks and selects linking data from the document understanding module.

4.3 Model of Information Structure for Information retrieval [2]

By analysis of information reterival process, we proposed a model of information structure for information retrieval (Fig. 1.)

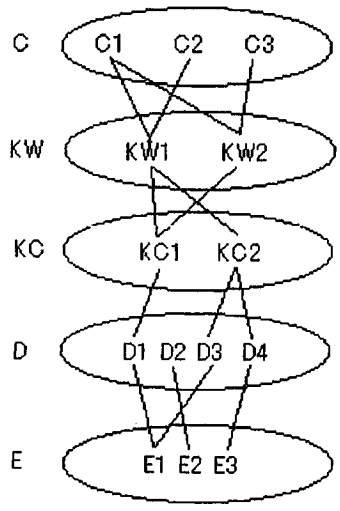

Figure 1. Model of Information structure

In this model, at first something happens, e.g., an social event or new technical invention. And they are described and reported by journalists or researcher in the form of newspaper article or scientific paper. Therefore, events (E) including invention and etc., are lined to documents (D) in the Fig.1. Next, in the document, there are some key words (KW) conveying important meanings, which corresponds to key concepts (key concepts, KC.) The lines between KW and KC shows this relationships (Note that KC layer is inserted between D layer and KW layer.) Key words are usually assigned to themes or categories. And key words are usually assigned to themes or categories, so KWs and Cs are line to each other. For more details, see the reference [2].

5. Evaluation of the Prototype System of ADAMAS-1

5.1 Materials of Evaluation

A set of text data provided by NACSIS was used to evaluate the prototype system of ADAMAS-1 in terms of a set of queries also provided by NACSIS.

5.2 Method

As ADAMAS-1 is under the process of implementing parts of the system, and does not have enough knowledge for advanced information retrieval, we adopted the conceptual terms in the information retrieval queries given by NACSIS as key concepts tentatively. Information retrieval was done with use of a set of key concepts.

5.3 Result

No rigid results are now available (we are sorry,) because we are still investigating the results of the evaluation.

5.4 Considerations

Through the evaluation with use of NACSIS data, the followings are clarified:

(1) ADAMAS-1 is not practical now, because this system is assumed that the system has enough system knowledge, but it is, for the presence, infeasible to construct it.

(2) Systems like ADAMAS-1 should have capabilities of acquiring system knowledge autonomously, because the knowledge is usually to huge to construct it manually.

Other problems to be coped with are now under investigating.

6. Conclusion

In this report, we presented our idea of a new advanced information retrieval system in some details and the evaluation results of the idea.

<<Acknowledgement>>

This project is partly supported by the Japan Society for the Promotion of science (Research project No. JSPS-RFTF96R15201.)

<<References>>

[1] Fujisaki et al. : "Man-machine dialogue system through spoken language," Research Report of the Research for the Future Project, pp.1-68(1998).

[2] Hiroyuki KAMEDA and Hiroya FUJISAKI: "Newspaper article information classification and retrieval system with use of layered sturcture of theme-key concept-key word," Journal of Information Processing of Japan, No.11, pp.1103-1111(1987).