Creating an Age Where Anyone Can Find the Information They Truly Need

NTCIR's Information Retrieval Ideal

With the spread of sites such as Google and Yahoo!,
using search engines to find information has become commonplace.
What is the optimal information for users?
We asked 3 researchers involved in NTCIR, a workshop-style project
whose objective is the continued development of information access technology.

What is NTCIR?

With the spread of computers and the Internet, hearing news reports of topics such as people playing shogi or chess against computers has become commonplace. We live now in an era where computers appear on game shows. A research and development project at IBM is currently working on a computer system which will participate as a contestant on the popular American quiz show "Jeopardy!". It contains a question response system that combines information retrieval and language processing technologies, and research is underway on question answering technologies that can also handle so-called "trick questions".

NTCIR is an internationally active, workshop style project whose objective is the advancement of these kinds of information access technologies. The expression "information access" was chosen because NTCIR's objective is a system for "supporting the users to create of new value from massive amounts of information"by retrieving relevant information for the users and supporting the users to utilize the information in the documents. Therefor NTCIR is researching information retrieval, technologies for supporting the users to utilize the information in documents such as question answering, summarization, opinion analysis, trend analysis, etc. and search appropriate questions.

The project, started in 1997, subsumes several research divisions related to information access technologies, and each division is operated by a group of researchers, who function as "organizers". The importance the project places on the workshop-style approach can also be seen in research division selection. The project's management does not unilaterally establish research divisions. Instead, researchers in related fields collect research proposals, and decisions are made after a committee considers them from the perspectives of their content, feasibility, international research trends and technological trends, and social value. Each activity cycle lasts a year and a half, with NTCIR-8 (its 8th cycle) currently underway. A single cycle's process is as follows. First, organizers propose research division objectives and evaluation methodology, discussing them with researchers who wish to participate, and determining final evaluation methods and data. After this, organizers distribute shared document datasets and query datasets. Participants use these datasets in their search experiments, thereby performing verification of the systems they have developed. These results are collected, evaluated by human assessors, and correct answers are created. In some cases, correct answer candidates, created in advance, can be used, in which case many participants evaluate and verify them, increasing the reliability and validity of the correct answer proposals. Last, the verification and evaluation results are gathered in the form of research papers, which report the achievements of the research, bringing the cycle to a close. The document data, query data, and correct answers are referred to collectively as a "test collection". These are, of course, repeatedly used by researchers participating in NTCIR for further research, but they are also made publicly available to non-NTCIR participants, so that a wider research community can test their methodologies against a validated standard. This contributed to the efficient advancement of research activities. Verifying the effectiveness of information access technologies requires, for experimentation, a large number of users and questions. However, during initial stages of research, where verification is needed every time a new idea emerges, it is difficult to gather a large number of users and perform long-term testing. By using test collections, the effectiveness of research ideas can be verified immediately, and repeatedly, through research lab experimentation, rapidly accelerating research progress.

A Half Century of Information Retrieval System Progress

NII's Professor Noriko Kando has been deeply involved in NTCIR activities since it was first established. "Computer system based information retrieval research first started in the 1950s," she said, providing the following overview of the history of the research.

Shortly after research began into information retrieval, it diverged into two streams: one focused on making it commercially viable, the other on theoretical research into search algorithms and the like. The searches

Noriko Kando
Professor,
Information and
Society Research Division, NII

Atsushi Fujii
Associate Professor,
Tokyo Institute of Technology

Koichi Takeda
Senior Manager,
Senior Technical Staff Member,
Information & Interaction,
IBM Research - Tokyo

performed on the commercial systems of the time were exact match searches, searching for exact matches of search queries. Most were simple searches that looked for matches in the titles or abstracts of papers. However, with improvements in hardware development technology came a corresponding dramatic increase in the amount of data handled, requiring a re-evaluation of search approaches. Specifically, searching for a single word would result in an enormous number of hits, while searching for multiple terms would result in excessive narrowing of search focus often producing no search results. In order to resolve this problem, the ball was passed to the other stream of research, that focused on the development of search algorithms. "While commercial systems used exact matches, their research used the best match approach. The objective of this method was to present information in response to user information needs in ranked order of relevance."

In order to put the results of search algorithm research into practice, comparative validation of technologies and methods using large scale test collections was essential. This was the start of research by experimental evaluation workshops. The first was TREC, established in the US in 1992, followed by Japan's NTCIR and then Europe's CLEF. Through the establishment of these research projects, continued research result technology transfer, and close-knit research exchange, evaluation workshops became sites for new and ever challenging research task proposals.

The Best Match Search Approach

What kind of search methodology is employed by the "best match" approach, which searches for relevant information? With the best match method, search systems rank search results in the order predicted to be most relevant to the user. It includes evaluations of the frequency with which words or strings contained in searched documents and questions appear, in what patterns they appear, document lengths, and the like, and in order to calculate their similarity, create mathematical models and "retrieval models" (vector space models, probabilistic models, language models, etc.). A variety of other heuristics are also used, such as web links and usage histories, including click logs. These are used to determine similarity between searched documents and queries, to calculate importance measures, and to rank search results. The best match method does not merely search for strings contained in the query, but for the information that matches the information needs and intentions behind the query.

The Uses of Information Retrieval Systems, Extending Beyond the Bounds of the Web

The term "information retrieval" might call to mind search engines such as Google and Yahoo!, but information retrieval technology extends beyond just the web. One of NTCIR's task organizers, Koichi Takeda, an IBM Japan employee who specializes in machine translation and text mining (*) research, says: "The majority of office information is unstructured (text, image, and other), which contrasts with structured information in a database. It is said that up to 30% of a white-collar worker's time is spent searching for and analyzing information. For improving these workers' productivity, we need to develop solutions that provide information in the way that users can easily access and manage, both quantitatively and qualitatively - that is, solutions that provide relevant information."

Wide-Ranging Research, from "Patents" to "Yahoo! Answers"

NTCIR started in 1997, and is now in its 8th cycle, gradually turning into an international research project,

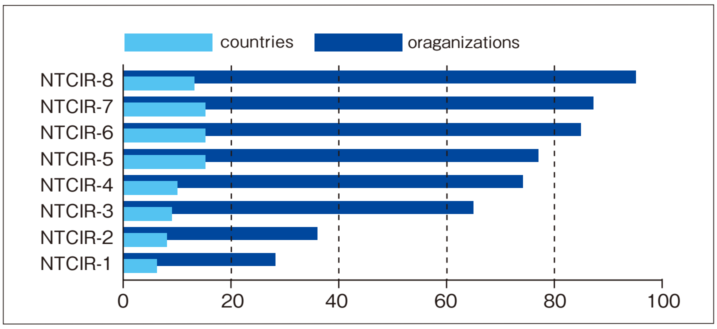

Figure 1 Trends in Countries and Organizations Participating in NTCIR

with participating researchers from 17 countries (figure 1). On the next page follows a few of NTCIR's notable research projects.

One of the fields NTCIR has researched since its formation is cross-lingual information retrieval, performing searches across multiple languages. One of NTCIR's groundbreaking researchers, who has participated in the project from its earliest days, Tokyo Institute of Technology's Associate Professor Atsushi Fujii, provides the following overview of the NTCIR's first cross-lingual information retrieval system. "Let me tell you about the cross-lingual information retrieval system developed for searching English academic papers with Japanese language queries. When performing searches with this system, there are two possible approaches: Japanese queries can be translated into English, or English papers can be translated into Japanese. The latter places a heavy burden on system processing, so generally the former is used. However, in order to obtain more accurate search results, the following approach can be used. Queries can be translated into English and searching performed. Then the top results, let's say the top 1000 results, are translated into Japanese, and a search is run again against the original Japanese query. This method came to be called "bi-directional", and is now the primary method used in cross-lingual information retrieval."

Patent information searching, started at NII in 2001, was one of the research fields that, internationally, NTCIR focused on from an early point. A number of issues were faced during research, but these were surmounted with the research assistance and backup of companies providing patent information services, and intellectual property information retrieval committee members from the Japan Intellectual Property Association. As a result, a variety of research was done on topics such as cross-lingual and cross-document type searching, and automatic "patent map" technical trend generation, and a considerable amount of the research has since been used in commercial applications.

Question answering, which extracts answers themselves from large volumes of documents, is another research division on which NTCIR has focused. Research so far diversed from working with simple factual questions such as "Who is the prime minister of Japan?", to dialog style questions with complex responses such as definitions and relationships, such as "What is information retrieval?", to answers such as "There is no answer" for unanswerable questions, and further still to cross-lingual question answering. Document searching, question answering, and summarization have developed as separate technical fields, but NTCIR believes that these technical fields will fuse in order to provide users with relevant information in an appropriate format.

Research underway on searching Yahoo! Answers is also quite unique. Yahoo! Answers is a web service where those with questions post them, and other users post answers, the best answers being marked as "Best Answer". In addition to user evaluations, NTCIR is researching a system for automatically determining the best answer. In order to increase objectivity, in addition to the best answer selected by the asker, Yahoo! Answers also had other evaluators select best answers. "Analyzing the answers selected as best answers, we saw certain trends emerging, such as comments indicating approbation of the questioner, or the insertion of URLs providing evidence for the answer. That means that evaluations differed even for the same answers based on how the responses were written. This shows that considerations regarding communication style and modes of expression were linked the type of information provision required by the asker," explains Professor Kando. These research results can contribute text retrieval an to service improvements for Yahoo! Answers, but can also be applied to a wider range of communications. For example, it would be reasonable to use such systems to automate responses to inquiries to company service desks. "Information such as that handled by Yahoo! Answers carries considerations of user privacy, making it difficult to use in research, but some of those involved in the startup of Yahoo! Answers are familiar with NTCIR's activities, and provided dataset in accordance with proper established procedures for maintaining user confidentiality."

The Allure of the Workshop-Style

NTCIR-7's research result presentation conference

NTCIR's research is supported by a large base of people. This is not limited to coordination with external organizations. Professor Kando explained the significance of coordination. "I believe that the greatest allure of NTCIR is its workshop-style, with people coming together to tackle research issues. When research is performed by a single organization, there is a limit to the number of ideas and testable approaches, and it is difficult to perform objective evaluations. During the year and a half of a project cycle, NTCIR establishes several roundtable meetings as opportunities for exchanging ideas. There is particular excitement among researchers concerning initial results reported at the meeting roundtables. Researchers are tackling the same issues, so there is a high degree of identification with details which cannot be fully encapsulated in papers, such as approaches to the research theme, and experimental know-how."

The workshop-style also presents significant advantages in terms of technological coordination. For example, take question answering systems. These can be broken down into multiple phases, such as information collection, analysis and extraction, and aggregation and presentation. In other words, component-based is possible for systems. This means that there are many cases where, instead of a single organization developing every function, individual organizations develop the modules in which they have particular strengths, combining them to form a superior system.

Future Developments in Information Access Technology Research

NTCIR-8 will ends this June, but what will future NTCIR activities bring? "The number of information retrieval researchers is low in Japan, compared with other countries, and I believe that reinforcement is needed," says Associate Professor Fujii, who provides direction to students himself. "As someone who works in the classroom, I hope to raise not only people who can design and develop search systems, but also people who can use test collections and perform appropriate system evaluations. Evaluating search systems is just as important, and as difficult, as designing and developing them. I also think that evaluating search systems and grading students share something in common. Test collections, and test questions administered to students, must be based on fair evaluation standards, and a wide variety of questions, materials used in solving them, and correct answers, with appropriate difficulty levels, must be created. Students and systems evaluated with these must be at a level at which they can perform practical tasks in the real world. My desire is to nurture students and systems which will contribute to actual society." Dr. Takeda, as a member of the business world, expressed his visions, saying, "I would anticipate that emerging applications of NTCIR research are going to be proposed to meet real world demands. So far, research has focused on individual components such as searching, translation, or text mining. By combining them in the future, more sophisticated information access solutions will be introduced to innovate office activities." Lastly, Professor Kando discussed two future objectives. "The first is exploratory search. Web search engines are used for searching for information regarding actualities, and when the user knows that there are answers, such as searching for NII maps or tomorrow's weather forecast. However, there are also a large number of cases involving taking an interactive search and learning approach, such as when it isn't clear what the user is looking for, when the topic is in the area which is not falimilar with the user, or when the goal of the search is unclear. I hope that this kind of interactive exploratory search and information utilization is made possible in order for the answers to which search systems lead users be brought even closer to what the users are looking for. For example, consider a hypothetical mother searching for a nursery school in which to enroll her child. Parents naturally want to send their children to good nursery schools, but for a first-time parent, I don't think they know what "good" aspects to look for. I think, in this kind of case, that these aspects must be provided as possible selections. Another example of people searching very vague topics would include, for example, high school students who wish to find out about "college entrance exams". If searching for "college entrance exam" produced university rankings by department, procedures for taking entrance exams for foreign universities, graphs showing the career tracks of graduates, and the like, it would be easier for said high school students to take the next step. I want to create systems that make it possible for users to take an exploratory approach to information searching, learning, and investigation, by presenting the perspectives necessary for searches, categorizing and aggregating search result information, and arranging it for presentation. This technology, by aggregating massive volumes of data, would support the creation of new value by users, pursuing the weaving of knowledge from information, the goal of the NII, from the standpoint of information access research. Interactive information access technology evaluation methodology must be established as a foundation for this research, and research into this methodology is particularly active on an international level.

The second objective is closely related to the first, being the functioning of NTCIR as a community in the true sense of the word. I hope that the opportunity presented by NTCIR is made full use of for researchers, as organizers and participants, to openly develop and advance their own research, and to nurture students and young researchers. I believe NTCIR's role is not to lead this self-motivated community, but to provide support for it. Doing so would increase the effectiveness of the workshop-style approach, resulting in coordination which makes optimal use of individual specializations, and the creation of even better systems."

This open research style is sure to generate new, heretofore unimagined technologies and value.

(Written by Takuya Kudo)

*Text mining: includes (1) information extraction from large volumes of text data - for identifying words, named entities (names of people, places, etc.), emotional expressions, "subject/object-predicate" expressions, and the like, and (2) analysis of their frequency, patterns, and relationships in order to acquire insights that is not immediately apparent, and to support document organization and report generation.