A Study of Integrate Framework and Research Environment Concerning Mandarin Chinese Full Text Information Retrieval

Yun-Long Huang[1], Eric S. H. Liao[2]

[1] National College of Physical Education and Sports, Assistant Professor of Sport Management (ylhuang@mail.ncpes.edu.tw)
[2] Academia Sinica, Research Assistant of Information Science(ericliao@gate.sinica.edu.tw)

Abstract

Full text information retrieval (IR) becomes the focus of interest in the area of interdisciplinary study. This paper attempts to clarity some basic questions concerning the full text information retrieval as well as the scope of theory and related research.

Concerning Mandarin Chinese IR is facing more basic difficulties than English IR because of research lag and language nature. Lack of an objective test collection for IR experiments is the fundamental issue for Mandarin Chinese IR. In this paper, based on the experience of test collection construction, we develop some criteria as research guidelines for corpus selection, document classification and indexing, query design and relevance judgment.

Finally, we propose a design of fundamental research environment including seven modules. There are test collection, full text database management system, automatic indexing, query processing interface, search engine, performance evaluation, and relevance feedback.

Keyword : Research Framework : Full Text Information Retrieval : Information seeking behavior : Relevance : Information Needs : Test Collection

1. Introduction

Full text information retrieval (IR) becomes the focus of interest in the area of interdisciplinary study. In facing the fast changing environment, new issues boost more complicated challenges. Concerning Mandarin Chinese IR is facing more basic difficulties than English IR because of research lag and language nature.

Lack of an objective test collection for IR experiments is the fundamental issue for Mandarin Chinese IR. There are still many critical issues have to be addressed which including information seeking behavior, fundamental research environment of IR research, IR system and theory, and more interdisciplinary study.

2. Research Question of Information Retrieval

Full text IR is the process of organizing documents by their features and retrieving relevant documents according to user's information need. Considering digital documents retrieval, integration of specific field-knowledge about the documents is quite essential in the retrieval course.

The basic question of IR can be: How does the system retrieve documents relevant to user's information need. To answer this question, users' subjective information needs and objective document contents must be addressed.

Language is the formal representation of concept. Characters, including ideogram and phonogram, are symbols representing language. Therefore, the main conceptualized question relating to IR is "what is the relationship between formality and content". Words are the basic, independent, and meaningful language unit. The meaning of a word is the representation of concept being thought. In other words, concept is the content of word, and word is the formality of concept.

A full text IR system should be a structured formal one. Contents and topics of all documents in the system have to be transformed into system dependent representation to be properly retrieved. Besides, users' information needs to be transformed into description of query, then into format capable of being processed by the system. To be short, the essence of IR is the relationship between "formality and content".

3. Integrated Research Framework of Information Retrieval

The main function of a research framework lies in defining research scope and leading research direction. Through such approach, people can accumulate and build theory paradigm. That's to say, a definite research framework will pave the way for future research.

In this article, the authors construct a referential framework of full text IR research by referring to "A Framework for Research in Computer-Based Management Information Systems" by Ives, Hamilton, and Davis [1], and "Relevance, Pertinence, and Information System Development" by Kemp [2].

Fig. 1 describes a research framework model by three variables: context, user, and system. By simplifying "A Framework for Research in Computer-Based Management Information Systems", the authors keep focus on user's information need and the concept of full text IR.

In the framework, several different research scopes can be found, including information seeking behavior, IR research experiment environment, and IR system and theory. As a specific research scope is set, minor issues can be defined. For example, in the scope of IR system and theory, document collection is the basis of automatic indexing, and query is the basis of people-machine interface, natural language query, and query expansion. If a research scope across multi-fields is set, much more complicated experiment design is required. For example, the issue of retrieval performance evaluation may span the three main areas in fig. 1, and the question may be more complicating.

If the research scope extends to information technology, such as web search, online database search, new space for theory development will emerge from its original single and simple issue. For example, intelligent agent on web must take both search engine and users' information need into account. Moreover, organization environment change will make the question more troublesome. Only by the support of basic research can progress be made.

Reviewing earlier research with the submitted framework, automatic system was the main focus. Planning, development, implementation, and operation of retrieval system, however, were not important relatively. By 1961, The SMART project, based on documents of Cranfield research, started to make the pace toward the development of theory of automatic index. Vector Space Model proposed by Salton, and used in the SMART project, is still a valuable theory.

From then on, interaction between library science and information science is getting intense. Research of Researchers of Information science focus on IR system and retrieval theory, such as automatic indexing, automatic document classification, clustering index, thesaurus auto-generation, relevance feedback, fuzzy query, AI learning, and natural language processing, while researchers of library science focus on the impact of information technology on knowledge management and information seeking behaviors.

The increase of new issues is in consequence of change in emerging technology. For example, cross-language IR, multimedia IR, intelligent IR, and network resource IR etceteras. Moreover, the natural language processing techniques introduce to IR also stimulate in advance research of information extraction and document summarization; and these new technology align with information seeking behavior research explore the new topics of information routing, information filtering, and personalize information service. [3]

The advanced research topics above reveal that not only library science and information science are involved, many more fields, such as linguistics, sociology, mass communication are also key players. Therefore, we present a research framework in search of the integrated approaches. Through such approach, people can accumulate and build theory paradigm.

Social-Cultural Environment (Multi-Lingual Environment)

Organizational Environment (Enterprise, Library, Data Center)

Technology Environment (Network Publish, Online Database, CD-ROM Title)

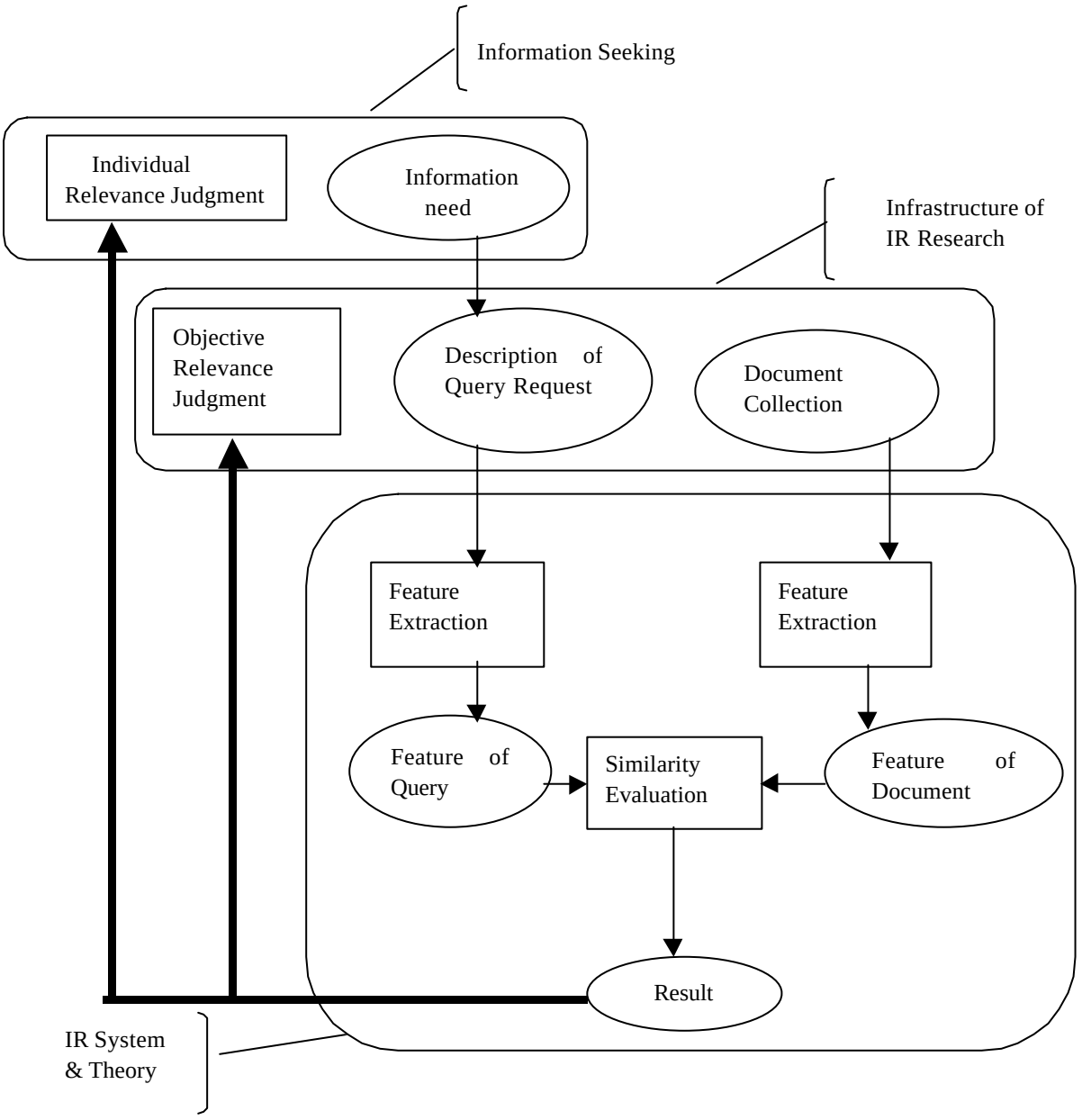

Fig. 1 Integrated Research Framework of IR

4. The Basic Concept of Integrated Research Framework

Based on the above discussion, two critical elements of full text IR research can be found: user information need and relevance judgment. In the following, these two elements are to be discussed.

4.1 Information Need

In the course of information retrieval, user's information seeking behavior is highly involved. It is a description of the activities ranging from user's awareness of his information need to satisfaction of his information need. "Information need is a psychological state associated with uncertainty, and with the desire to know and unknown."[7] "It is not directly observable…but it has a definite existence in the mind of the user at least and so it is useful to have a term by which one may refer to it."[4] · In order to input a query to a system, the query must be transformed into a request, which is acceptable by the system. Both query and request can't adequately express human's information need. To some extent, this is a narrow definition of information need.

Taylor distinguished four stages of information need: the visceral need, the conscious need, the formalized need, the compromised need [5]. In the first stage, by the time when a person start to be in need of certain information, it may be vague, or even imperceptible. In the second stage, as the need getting stronger, its existence and definition is clarified gradually. In the third stage, the person can describe his perceived information need with certain system of symbols. In the final stage, the person may modify his description about information need to fit the specification of the IR system.

In conclusion of the above discussion, user's information needs can present in three different types. The one is only existence in the mind of the user. The second is clarified with certain system of symbols. The third is compromised with the specification of the IR system. Since the first present type is difficult to measurement, IR research should be start up from the second type.

4.2 Relevance Judgment

Relevance judgment is the foundation of IR performance evaluation. In the field of IR, relevance means the relatedness of a document and a query. For each retrieved document, the relatedness between the document and the query is judged. In this way, retrieval performance can be measured. Then, comparison among different retrieval strategies or IR systems is possible.

Schamber et al. remind that the need for a thorough definition is at least three reasons [6] :

(1) "Relevance is the measure of retrieval performance of all information systems, including full-text, multimedia, question-answering, database management, and knowledge-based systems. Increasingly complex systems are being developed that promise to serve users more effectively than ever. It is inevitable that these new systems, like systems since the library card catalog, will be evaluated (Explicitly or implicitly)on the basis of human relevance judgment."

(2) "Among current development are information retrieval systems that actually employ users' relevance judgments in the course of their operating processes. For example, relevance feedback mechanism, in which users' relevance judgments are used to modify list of documents as the search progresses, make users an integral part of the system. In such systems, relevance is no longer a reactive concept, to be used primarily in evaluation, but an active concept vital to the functioning of the system itself. However, without an understanding of what relevance means to user, it seems difficult to imagine how a system can retrieve relevant information for users. "

(3) "Information scientists must finally establish a full theoretical and empirical understanding or definition of relevance as a fundamental concept, so that the discipline can move on to other matters."

As Saracevic(1975)indicate S. C. Bradford was the first one to use the term relevant in the context of 1930s that it is used today in information science. Then the more and more research on relevant from different points of view. [7] · Though many researchers delivered theory and definition of relevance, including topic relevance, logical relevance, contingency relevance and psychological relevance, a universal, generally accepted definition of relevance is not available. Currently, objective topic relevance is most widely used. [8]

According to definition of topic relevance by Cuadra and Katter: “Relevance is the correspondence in context between an information requirement statement and an article, i.e. the extent to which the article covers material that is appropriate to the requirement statement.”[4] Topic relevance deals with the relatedness among topics. If overlap between the topic of a document and that of a query exists, there is relevance.

In objective relevance, individual’s specific judgment is discarded. Instead, a general, normal point of view is adopted in relevance judgment. On the other hand, subjective relevance values individual’s information need and emphasizes that only the originator of the information need may decide whether a document is relevant.

5. The Experience of Test Collection Construction in English

There is over 30 years history on information retrieval. Research started with experiments in indexing languages, such as the Cranfield I test in 1960s. Users’ information needs assessment and relevance judgments are always hard work in research design. Following will introduce the development experience of Cystic Fibrosis (CF) document collection. In table 1 shows at least thirteen traditional test collection, with fewer than 30,000 documents, and four TREC collections. In such cases we can find the difficulty in development processes and resources requirement. [9]

The CF document collection consists of 1,239 papers published in the years 1974 through 1979 and indexed with the term CYSTIC FIBROSIS in the National Library of Medicine’s MEDLINE file. The CF database includes 100 queries with three sets of exhaustive relevance evaluations from subject experts. The exhaustive, full-text relevance evaluation means that subject experts judge a document to be topically related to the query and assigned the document a relevance score. 14 subject experts formulated the 100 queries. Among them, 9 subject experts execute exhaustive relevance evaluation base on their domain knowledge, 4 postdoctoral researchers execute exhaustive relevance evaluation base on their domain knowledge, finally one experienced medical bibliographer execute exhaustive relevance evaluation.

Table 1 IR Test Collection Statistics

Test collection	Number of documents	Number of queries	Average number of relevant documents
TIME	425	83	3.9
KEEN	800	63	14.9
MED	1,033	30	23.2
CF	1,239	100	31.9
CRAN	1,400	225	8.2
CISI	1,460	76	41
HARDING	2,472	65	22.6
EVANS	2,542	39	23.1
CACM	3,204	52	15.3
LISA	6,004	35	10.8
NPL	11,429	93	22.4
INSPEC	12,684	77	33
UKCIS	27,361	182	58.9
TREC-3-WSJ	173,252	50	78.3
TREC-2-WSJ	173,256	50	91.1
TREC-3-full	741,856	50	196.1
TREC-2-full	742,611	50	232.9

In recently, the Text Retrieval Conference (TREC) developed the most famous test collection. The overall goal of the TREC is providing a very large test collection, and encouraging interaction with other groups in a friendly evaluation forum, then a new thrust in information retrieval will occur.

According to table 1, the scale of traditional test collection is relative small. The sources of TREC test collections

include Wall Street Journal (1986- 1992), AP Newswire (1988-1990), the Federal Register (1988-1989), Computer Select disks (Ziff-Davis Publishing), publication of Department of Energy, San Jose Mercury News (1991), and U.S. Patents(1993).[10]

In such large collection, relevance judgment will be critical challenge. For example, in TREC-2-full 742,611 collection, for each topic, resulting in over 74 million judgments. It is clearly impossible. Therefore TREC construct the "pooling method". TREC make relevance judgments on the sample of documents selected by the various participating systems. It was the recommended method in 1975 proposal to the British Library to build a very large test collection. [11]

6. Key Issues Relating to Test Collection Construction

To construct an objective platform of Mandarin Chinese full-text IR experiment, variables, such as corpus type, document classification and index, description of user information need, relevance judgment, must be controlled. The following is some general principles.

6.1 Selection of Document Collection

Different type of document collection may influence IR performance. In selection of proper document collection for IR experiment, several questions must be considered:

(1) The type of document content: Documents may be classic or of colloquialism. Documents may contain more dialect and hard to understand, or plain enough for most people to understand. Documents may concentrate on some fields, or distributed relatively equally among many fields. Subject concentration in specific domain knowledge is additional consideration. It is an important factor in evaluation of system performance in different test collections.
(2) The language of documents: Documents may be composed purely by Chinese characters or contain multi-lingual words.
(3) Size of the collection: Large collection is more persuasive, but less easy to build.
(4) The internal encoding scheme of text file: The current Big-5 code accommodates less than 6000 commonly used Chinese characters. Other less used characters have to be created or defined individually.
(5) The media of the collection: If only paper documents are available, the process of digital data transformation is needed.

6.2 Document Classification and Index

Considering the validity of research design, an ideal document collection should be well organized, and provide scientific knowledge of classification, subject heading analysis, document indexing, and thesaurus. Professional librarians should execute such tasks.

6.3 Description of Information Need and Query Design.

If previous real user queries are recorded, then researchers can use them directly or after proper modification. Besides, those required queries are based on real information need and more close to real information seeking behavior. If none of them exists, then researchers have to re-invent a set of query for IR experiment.

6.4 Relevance Judgment based on Query and Documents

Relevance judgment is quite tedious, but essential. Except definition of relevance, that how to produce a correct, consensus, and thorough set of relevance judgment, under the restriction of resources is quite a problem. Sophisticated and elaborated judgment may suffer difficulties as collection size grows.

7. Fundamental Research Environment Planning of Information Retrieval

In this paper, based on previous research, [12] the authors proposed a basic model for IR research (Fig.2). The model includes seven modules: corpus, full-text database, automatic indexing, query processing, search engine, performance evaluation, and relevance feedback.

Automatic indexing: controls building document feature according to document content. If term segmentation is

required, it must be performed before indexing.

Search engine: compares the features of documents and query and selects those relevant documents.

Performance evaluation: computes the effectiveness of the system and presents the results visually.

Query processing and relevance feedback: provide man-machine interaction interface. Based on the retrieval result, a user can submit his subjective relevance judgment to the system for retrieval refinement.

Full-text database: ease data maintenance and increase access efficiency.

The model described above, is just a basic planning, and only defines the main function of modules and the relationships between them. Implementation of the model relies on future project for a better IR experiment environment.

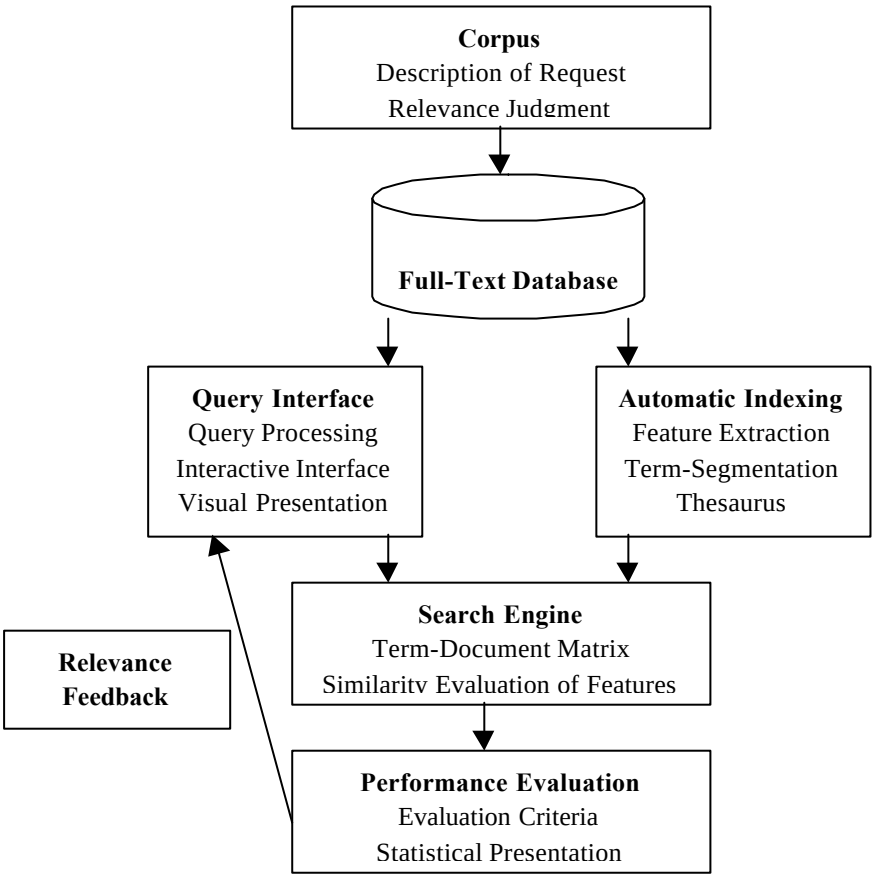

Fig. 2 Fundamental Research Environment Planning of IR Research

References

[1] B. Ives, S. Hamilton, and G. B. Davis, "A Framework for Research in Computer-Based Management Information Systems," Management Science Vol. 26, No. 9, pp910-934, 1980

[2] D. A. Kemp, "Relevance, Pertinence, and Information System Development" Information Storage & Retrieval Vol.10, No.2, p38, 1974.

[3] Y. L. Huang, "A Study of Research Framework and Critical Issues Concerning Mandarin Chinese Full Text Information Retrieval", University Library Quarterly, Vol. 2, No. 3, pp4-26, 1998.

[4] W. S. Cooper, "A Definition of Relevance for Information Retrieval" Information Storage & Retrieval, Vol. 7, p21, 1971.

[5] R. S. Taylor, "Question Negotiation and Information Seeking in Libraries," College and Research Libraries, pp182-183, 1968.

[6] L. Schamber, M. B. Eisenberg, & M. S. Nilan, "A Re-examination of Relevance: Toward a Dynamic, Situational Definition" Information Processing & Management Vol. 26, p756, 1990.

[7] T. Saracevic, "Relevance: a Review of and a Framework for the Thinking on the Notion in Information Science," Journal of the American Society for Information Science Vol. 26, 321-343, 1975.

[8] Chung-Chiao Lu, “The Developments and Trends of Relevance for Information Retrieval,” <http://www.lib.nccu.edu.tw/mag/16/16-3a.htm>,(Accessed April 22, 1998).

[9] W. M. Shaw Jr, Robert Burgin, and Patrick Howell, "Performance Standards and Evaluations in IR test Collections: Vector-Space and Other Retrieval Models," Information Processing & Management Vol. 33, p19, 1997.

[10] D. Harman, "Overview of the Fourth Text Retrieval Conference (TREC-4)," in Proceedings of The Fourth Text Retrieval Conference (TREC-4), ed. By D. K. Harman(NIST Special Publication 500-236), pp1-24, 1996.

[11] D. Harman, "Overview of the First Text Retrieval Conference (TREC-1)," in Proceedings of The First Text Retrieval Conference (TREC-1), ed. D. K. Harman(NIST Special Publication 500-207), pp1-20, 1993.

[12] Eric S. H. Liao, “Research of Planning and Construction of Test Collection for Mandarin Chinese Full-Text Retrieval,” Unpublished Master Thesis, Graduate Institute of Business Administration, National Taiwan University, June, 1998.