情報学基礎 51－5
自然言語処理 127－5
（1998. 9. 17）

ユーザの情報利用目的に基づく検索システム

全裕里, 石間衛, 藤井敦, 石川徹也
図書館情報大学
{youlee,ishima,fujii,ishikawa}@ulis.ac.jp

キーワードを質問に用いる検索システムは、ユーザの情報要求とは関係のない文書を検索してしまうことがある。その結果、ユーザは大量の検索結果から自分が欲しい文書を探さなければならないことが多い。この傾向は、ユーザが「何かについて単に知りたい」というような漠然とした要求を持っている場合に顕著になる。そこで我々は、ユーザの情報利用目的に基づく検索法を提案する。本手法は、検索質問中に含まれる動詞を手掛かりにして情報の利用目的を抽出し、その目的に関連する表現を含む文書を検索する。予備実験の結果、本手法はキーワードに基づくシステムの検索性能を向上させることを確認した。

A Utility-Based Information Retrieval System for User Information Usage

Youlee CHUN, Mamoru ISHIMA,
Atsushi FUJII, Tetsuya ISHIKAWA
University of Library and Information Science
{youlee,ishima,fujii,ishikawa}@ulis.ac.jp

Conventional keyword-based information retrieval techniques, which are currently used for a number of search engines in the Internet, still find it difficult to reject noisy documents. Consequently users usually end up with browsing unreasonably enormous amount of retrieved documents. This problem is especially salient when users have vague requirements, such as “simply wants to know about something”. To resolve this problem, we target users who have explicit purposes for information usage. To realize this notion, we use query verbs as strong clues for user information usage and retrieve documents containing verbs associated with those usage types. Our preliminary experiments showed that our system outperformed keyword-based baseline systems.

1 Introduction

Given the growing number of machine-readable documents accessible via computer networks (the Internet, for example), information retrieval (IR) is a crucial task to provide information precisely related to user requirements. Existing information retrieval systems, most of which can be categorized into statistical approaches [4, 15], retrieve documents associated with words contained in user queries, and sort retrieved documents in descending order based on their importance degree[1]. However, most IR systems still find it difficult to exclude noisy (irrelevant) documents, and consequently users end up with browsing unreasonably enormous amount of retrieved documents. For one thing, most users cannot computationally formulate their real interests (motivations), and therefore they usually use a small number of simple key words as a query. Let us take a user who wants to buy a computer, as an example. This user may use the term *computer* as a query, and an IR system provides any information associated with computer. In fact, one can access more than 26 million documents containing the term *computer*, via the "Altavista" (keyword-based) search engine[2] on the Internet. This example suggests that we need to improve the precision of IR systems by enhancing user queries.

One solution would be "cognitive approaches", which model user interests into well-formulated queries prior to the retrieval process [3, 14]. These approaches are expected to be effective especially when a user has a vague motivation, i.e. simply wants to know about something.

On the other hand, user motivations can be well-represented by even simple natural language, given an explicit purpose for information usage. Hereafter we shall call this purpose "utility". This is exactly the case where we aim at in this paper. Let us reconsider the previous example, that is, a user who "wants to buy a computer". In this example, the term *buy* can be a strong clue along with *computer*. Given these clues, one may easily notice that relevant documents should be associated with the commercial of computers, not lecture of computer science technology. However, since no document for a computer sale literally contains "want computer(s)", we need to formulate the appropriate query. Intuitively speaking, we transform the query "want computer(s)" into "sell computer(s)" or "produce computer(s)", using a predefined verb dictionary [8]. By this, irrelevant documents are expected to be excluded from candidate documents, and thus the user can gain precise information.

Our system can also be characterized with the introduction of natural language processing (NLP) techniques, that is, morphological/syntactic analyses on both queries and documents. Unlike English, for a number of languages including Japanese, lexical segmentation is poignant along with part-of-speech tagging. Besides this, syntactic analysis is required to identify predicate-argument structures, such as between *want*(verb) and *computer*(accusative case noun). It should be noted that while our current implementation is targeting Japanese IR systems, our notion, namely 'utility-based' approach, is expected to be generally applicable to other languages.

Section 2 describes the design of our utility-based IR system, and Section 3 elaborates on our retrieval algorithm. In Section 4, we evaluate the effectivity of our system[3]. Before conclusion, we discuss future direction in Section 5.

2 Overall Design

Figure 1 depicts the overall design of our 'utility-based' information retrieval (UBIR) system, in which "database" refers to a document collection. Let us briefly explain the retrieval process based on this figure. First, "parser" conducts on morphological and syntactic analyses on sentences contained in the database. In the case of Japanese, morphological analysis involves lexical segmentation and part-of-speech tagging. Then, syntactic analysis identifies original forms of conjugated verbs and syntactic relations, amongst whom predicate-argument structures are extracted by "pred-arg extractor". To sum up, the database is compiled into "analyzed database" prior to the retrieval process. Thereafter, given a query comprising a simple natural language sentence, the parser and pred-arg extractor extract a predicate-argument structure ("pred-arg") from the query, as performed on the database sentence.

Although a number of Japanese parsers (morphological/syntactic analyzers, such as QJP [7] and KNP [10]) are proposed, we developed another parser from scratch because existing parsers do not necessarily target the application to IR systems. For example, inappropriate definition of lexical units (e.g. compound nouns) potentially de-

[1] In practice, documents with diminished importance can optionally be discarded from retrieval results.

[2] `http://www.altavista.digital.com/`

[3] We used BMIR-J2, a test collection for evaluation of information retrieval systems, based on the Mainichi Shimbun CD-ROM'94 data collection. BMIR-J2 was constructed by the SIG Database Systems of the Information Processing Society of Japan, in collaboration with the Real World Computing Partnership.

grades the performance of IR systems. However, our parsing methodology is beyond the scope of this paper and we do not further discuss this issue here.

Finally, "IR engine" retrieves relevant documents from the analyzed database, by the use of "thesaurus" and "verb dictionary". We use the thesaurus to expand query terms. Note that while we developed a thesaurus targeting our UBIR system, a number of machine-readable thesauri (such as Roget's thesaurus [2] or WordNet [12] in the case of English, and *Bunruigoihyo* [13] or EDR [6] in the case of Japanese) are fundamentally applicable. The verb dictionary lists verbs used for typical information requirements, and more than one verb associated with each corresponding requirement. In the case of the query example above ("want to buy computer"), the dictionary provides verbs, such as "to sell" and "to produce" for the query verb "to buy".

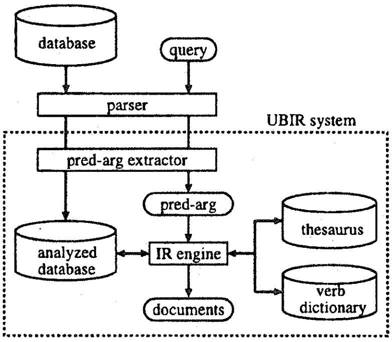

Figure 1: The overall design of the utility-based IR system

3 Retrieval Algorithm

In this section, we elaborate on our algorithm focusing mainly on the portion enclosed within the dashed region of Figure 1.

Let us take a Japanese query "映画が見たい (*eiga ga mitai*; want to see a movie)" as an example, to illustrate our retrieval algorithm. First (subsequently the parsing process), we extract the predicate-argument structure(s) from the parsing result of the query. Figure 2 shows the parsing result for the example query, in which each line corresponds to lexical units (morphemes). The first column denotes the IDs of each morpheme. The first and second columns show the syntactic dependencies between corresponding morphemes. The third column denotes Japanese morphemes (the Romanized characters and typical English translations are given in parentheses), and the fourth column denotes their parts-of-speech. Note that in the case of the example query comprising a simple sentence, the extraction of a predicate-argument structure is trivial. However, in order to accept longer queries comprising complex sentences, both parsing and pred-arg extraction are crucial.

Second, we expand the noun "映画 (*eiga*; movie)" by way of the thesaurus, as performed in so-called "query expansion" technique. In this case, one may notice that the thesaurus provides similar words to *movie*, such as *film* or *cinema*, and movie titles available in the database[4].

Third, we use the verb dictionary to transform verbs contained in the query. The verb dictionary lists (a) verbs used for utility-based queries, and (b) more than one verb associated with each corresponding query verbs, as shown in Figure 3. For example, the verb "見たい (*mitai*; want to see)" in the above query derives verbs like "公開する (*koukaisuru*; to release)" and "上映する (*joueisuru*; to run)".

Finally, we formulate new query using any possible combination between expanded nouns and transformed verbs (resulting from the previous processes), and retrieve documents from the analyzed database. In practice, we have two choices for this purpose: whether or not to contain postpositions in the query. On the one hand, postpositions, which specifies case role of nouns, are expected to reject irrelevant documents, and thus attribute to the high precision. On the other hand, as with most NLP systems, this approach can be fragile and potentially sacrifice the recall. We currently employ the second technique, that is, queries are in the form of "verb & noun", without postpositions (We will further discuss this dilemma in Section 4.2).

4 Evaluation

4.1 Experimentation

To evaluate the effectivity of our UBIR system, we used the Japanese benchmark collection BMIR-J2 [9]. This collection consists of 60 queries and 5080 articles (economics and engineering fields) collected from "Mainichi Shimbun" newspaper [11][5]. Each query corresponds to the 5080 articles, based on one of three ranks of relevance, i.e. *topically* relevant (rank A), *partially* relevant (rank B) and irrelevant (rank C).

[4] Since automatic identification of movie titles is currently difficult, we *manually* collected them from the database. This issue needs to be further explored.

[5] Practically speaking, the BMIR-J2 collection provides only article IDs, which corresponds to articles in Mainichi Shimbun newspaper CD-ROM'94. Users must get a copy of the CD-ROM themselves.

1	2	映画 (*eiga*; movie)	noun
2	3	が (*ga*)	postposition
3	–	見たい (*mitai*; want to see)	verb

Figure 2: Parsing result for the query "want to see a movie"

買いたい (*kaitai*; want to buy)	販売する, (*hanbaisuru*; to sell)	商品化する (*shouhinkasuru*; to commercialize)
就職したい (*shuushokushitai*; want to enter)	採用する, (*saiyousuru*; to employ)	募集する (*boshusuru*; to recruit)
見たい (*mitai*; want to see)	公開する, (*koukaisuru*; to release)	上映する (*joueisuru*; to run)
⋮	⋮	

Figure 3: A fragment of the verb dictionary

Our evaluation methodology proceeds as below. First, we produced new queries because the BMIR-J2 collection does not contain queries which literally represent the user utility. For this, we manually searched the original 60 queries for those potentially applicable to our evaluation. Then, we enhanced main verbs to identified queries referring to the relevant articles defined in the BMIR-J2 collection. For example, we modify the query "wants to know about movies" to "wants to *see* a movie" when relevant articles are mostly associated with newly released movies. As a result, we produced five queries as shown in Table 1. For the sake of enhanced readability, we list the English translations only for main verbs and accusative case nouns in the column of "query", from which one may notice that each of 5 verbs represents the purpose for the information usage.

Second, we identified relevant articles for each *new* query. To do this, we investigated only (topically/partially) relevant articles defined in the BMIR-J2 collection, because (a) browsing whole 5080 articles is an overwhelming task, and (b) our preliminary observation showed that the original rank C articles can virtually be judged irrelevant for our evaluation as well. In Table 1, the column of "#relevant" denotes the number of the BMIR-J2 A/B articles identified relevant for the new queries.

Finally, we compared our system with baseline systems in terms of the conventional precision/recall trade-off. As with most experiments for IR systems, we defined precision and recall as in Equation 1.

$$
\begin{aligned}
\text{precision} &= \frac{\#\text{ of relevant articles retrieved}}{\#\text{ of retrieved articles}} \\
\text{recall} &= \frac{\#\text{ of relevant articles retrieved}}{\#\text{ of relevant articles}}
\end{aligned}
\qquad (1)
$$

The baseline systems used in this experiment systematically retrieve articles defined "(topically/partially) relevant" in the BMIR-J2 collection, for each query. Practically speaking, the one system retrieves only rank A articles, while the other retrieves both A and B articles. According to the BMIR-J2 policy, (a) each query has to be the central theme in rank A articles, and (b) each query is partially associated with rank B articles. In other words, the first baseline system, in a way, simulates an intelligent IR system combined with thematic analysis, while the second baseline system uses a simple keyword matching technique. In Table 1, the column of "precision/recall" denotes the performance of three different systems (two baseline systems and the UBIR system), from which one can see that the UBIR system is superior to the baseline systems for any query. In addition, the UBIR system drastically improved on the precision of the baseline systems, sustaining the recall: only irrelevant articles can be successfully excluded by way of our methodology.

4.2 Error Analysis

The evaluation in Section 4.1 showed the effectivity of our UBIR system. However, the performance was still relatively unsatisfactory for Q_3 and Q_4 (see Table 1, for detail), in which our system retrieved an *irrelevant* article which contains the following sentence:

(1) 米国が減税について日本を批判する.
beikoku-wa genzei-nitsuite nihon-wo hihan-suru.
(The U.S. criticizes Japan for the reduction of tax.)

In this sentence, "日本 (*nihon*; Japan)" and "批判する (*hihan-suru*; to criticize)" represents the phrase *criticizes Japan*. In fact, the article containing sentence (1) focuses on the attitude of the Japanese

Table 1: Five queries used and precision/recall of different systems

	query		#relevant		precision/recall (%)		
	verb	noun	A	B	only A	A+B	UBIR
Q_1	enter	big 3 Japanese airplane companies	1	0	5.56/100	3.45/100	100/100
Q_2	buy	LCD	3	0	44.4/100	18.2/100	100/100
Q_3	criticize	reduction of tax	3	2	12.5/60.0	1.66/100	75.0/60.0
Q_4	criticize	reduction of consumption tax	3	1	19.0/75.0	2.47/100	66.7/50.0
Q_5	see	movie	1	0	14.3/100	4.76/100	100/100
average	—	—	5	0.8	11.6/78.6	2.80/100	83.3/71.4

government, not the reduction of tax. However, this article was considered relevant because (a) our system does not rely on case information, and (b) "減税 (*genzei*; reduction of tax)" and "批判する (*hihansuru*; to criticize)" in sentence (1) correspond to query terms, i.e. *reduction of tax* and *criticize*, respectively. One may argue that introduction of case information is expected to resolve this problem. On the other hand, this requires us a number of additional considerations. For example, in Japanese, case markers can be omitted or topicalized (i.e. marked with postposition *wa*), an issue which our framework does not currently consider. In addition, relative clauses, where the surface case markers of the head noun is omitted, pose a similar problem. To resolve this problem, the analysis of deeper case level (not surface level) is needed, which still remains as a difficult issue in NLP research.

5 Future Direction

In this section, let us discuss future direction.

First, since information retrieval is never one-pass process, we need to facilitate the iterative retrieval, as performed in "relevance feedback" [4] and "pseudo-relevance feedback" [1]. The basic idea of the relevance feedback is to enhance query terms based on documents judged relevant by users. While the relevance feedback method requires user intervention, the pseudo-relevance feedback method *automatically* enhances queries based on documents ranked with great relevance degree in the previous retrieval. We note that these concepts can be combined with our framework. One possible implementation would be to enhance entries in the verb dictionary, based on top-ranked documents. It should be noted that iterative IR systems require different evaluation criteria, along with the precision/recall trade-off. One effective criterion is to compare the time taken to retrieve desired documents [16].

Second, we note that the structure of "analyzed database" (see Figure 1), which is currently merely a set of parsed sentence, is needed to be more sophisticated. Ikeda et al. [5] proposed a method to to classify documents based on 5W1H (who, when, where, when, why and how) information, aiming at Japanese text navigation. This method is expected to facilitate easier information access in conjunction with our system, because documents targeted in our system (such as those about newly released products and movies) can be generally categorized based on the 5W1H axes.

Finally, while we used five queries and 5080 articles in our experimentation (see Section 4.1), larger-sized test collections (especially a larger-sized query set) are invaluable for the further evaluation. Existing methods to built test collections can be classified into the following three approaches. In the first approach, queries and documents are collected *independently*, and then human experts determine relevant documents for each query (Sheridan et al. [16] called this the "naive method"). As can be imagined, the naive method is generally time-consuming to built large-sized test collections[6]. To counter this problem, in the pooling method, a number of different IR systems *pool* retrieval results for each predefined query. Consequently, the number of documents to be investigated is expected to be reduced. However, this method is not reasonable for our purpose because there is few utility-based IR systems exist, to the best of our knowledge. Sheridan et al. [16] recently proposed namely the "seeding method" as the third approach. In this method, they first select documents which is (somehow) unique in the collection, as "seed" documents. Thereafter, they produce queries for each seed documents referring their lead paragraphs. This method can drastically reduce the human overhead, because unique documents, such as those about the "meltdown in Chernobly" and "bombing in Okulahoma City", are generally published in a certain limited period. In addition, as Sheridan et al. [16] identified, the seeding method is effective in the case where users have desired results already in mind. We note that this notion can also be applied to our

[6] The BMIR-J2 collection was built using the naive method, and thus the document size is smaller than those used in the Text Retrieval Conferences (TRECs).

utility-based IR system, and therefore our test collection is expected to be enhanced with minimal cost.

6 Conclusion

In this paper, we proposed an IR system based on user information usage. While conventional keyword-based retrieval systems achieved high recall, these systems still find it difficult to reject noisy information. To resolve this problem, we target users who have explicit motivation for information usage (utility), and used verbs contained in user queries as the clue for utility. To do this, we first extract predicate-argument structures (the tuple consisting of main verbs and their object nouns) from both a document collection and a given query, through NLP techniques. We then use the thesaurus and verb dictionary to enhance the query, which is used for the retrieval. We also evaluated our system by way of experiments, in which our system outperformed two baseline systems in terms of both the precision/recall trade-off. Finally, we discussed future direction: (a) adopting iterative retrieval process, (b) adopting more sophisticated information classification based on the 5W1H axes and (c) building larger-sized test collection.

References

[1] Chris Buckley, Gerard Salton, James Allan, and Amit Singhal. Automatic query expansion using SMART: TREC 3. In *The 4th Text Retrieval Evaluation Conference (TREC-4)*, 1995.

[2] Robert L. Chapman. *Roget's International Thesaurus (Fourth Edition).* Harper and Row, 1984.

[3] David Ellis. *New horizons in information retrieval.* Library Association Publishing, 1990.

[4] William B. Frakes and Ricardo Baeza-Yates. *Information Retrieval: Data Structure & Algorithms.* PTR Prentice-Hall, 1992.

[5] Takahiro Ikeda, Akitoshi Okumura, and Kenji Satoh. Information classification and navigation based on 5W1H of the target information. In *Proceedings of the JSPS-HITACHI Workshop on New Challenges in Natural Language Processing and its Application*, pp. 69–74, 1998.

[6] Japan Electronic Dictionary Research Institute. EDR electronic dictionary technical guide, 1995. (In Japanese).

[7] Masayuki Kameda. A portable & quick Japanese parser : QJP. In *Proceedings of the 16th International Conference on Computational Linguistics*, pp. 616–621, 1996.

[8] Yukio Kishimoto, Miyuki Sunouchi, Yasuhiro Tsukada, Shigeru Chiba, and Tetsuya Ishikawa. Information retrieval system based on text and query analysis. *Transactions of Information Processing Society of Japan*, Vol. 35, No. 5, pp. 908–916, 1994. (In Japanese).

[9] Tsuyoshi Kitani, Yasushi Ogawa, Tetsuya Ishikawa, Haruo Kimoto, Hidekazu Nakawatase, Ikuo Keshi, Jun Toyoura, Toshikazu Fukushima, Kunio Matsui, Yoshihiro Ueda, Tetsuya Sakai, Takenobu Tokunaga, Hiroshi Tsuruoka, and Teru Agata. Lessons from BMIR-J2: A test collection for Japanese IR systems. In *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.

[10] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, Vol. 20, No. 4, pp. 507–534, 1994.

[11] Mainichi Shimbun. Mainichi shimbun CD-ROM '94, 1994. (In Japanese).

[12] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine Miller, and Randee Tengi. Five papers on WordNet. Technical Report CLS-Rep-43, Cognitive Science Laboratory, Princeton University, 1993.

[13] National Language Research Institute. *Bunruigoihyo.* Shuei publisher, 1964. (In Japanese).

[14] R. N. Oddy. Information retrieval through man-machine dialogue. *Journal of Documentation*, Vol. 33, pp. 1–14, 1977.

[15] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[16] Páraic Sheridan, Jean Paul Ballerini, and Peter Schäuble. Building a large multilingual test collection from comparable news documents. In *ACM SIGIR Workshop on Cross-Linguistic Information Retrieval*, 1996.