PAPER Special Section on Document Image Understanding and Digital Documents

Document Image Retrieval for QA Systems Based on the Density Distributions of Successive Terms

Koichi KISE[†a)], ***Member***, **Shota FUKUSHIMA**[††*], ***and*** **Keinosuke MATSUMOTO**[†], ***Nonmembers***

SUMMARY Question answering (QA) is the task of retrieving an answer in response to a question by analyzing documents. Although most of the efforts in developing QA systems are devoted to dealing with electronic text, we consider it is also necessary to develop systems for document images. In this paper, we propose a method of document image retrieval for such QA systems. Since the task is not to retrieve all relevant documents but to find the answer somewhere in documents, retrieval should be precision oriented. The main contribution of this paper is to propose a method of improving precision of document image retrieval by taking into account the co-occurrence of successive terms in a question. The indexing scheme is based on two-dimensional distributions of terms and the weight of co-occurrence is measured by calculating the density distributions of terms. The proposed method was tested by using 1253 pages of documents about the major league baseball with 20 questions and found that it is superior to the baseline method proposed by the authors.

key words: *document image retrieval, density distribution, precision, question-answering*

1. Introduction

Question answering (QA) is the task of retrieving *answers* rather than documents in response to a question with an emphasis on functioning in unrestricted domains [1]. Since it enables us to realize a more natural mean of "information retrieval" as compared to the keyword-based retrieval of documents, it attracts a great deal of attention in recent years. Much effort has been made including TREC conferences [2], as well as application to the Web [3]. In addition, some research groups have started offering services of QA systems to the public [4], [5].

Question answering has been studied in the field of information retrieval and thus most of the existing QA systems work only on electronic text. But is it enough for us to deal only with electronic text? We consider that it is not sufficient because at least of the following two reasons. First, we have already had a huge amount of document images in various databases and digital libraries. For example, the magazine "Commȯf the ACM" in the ACM digital library [6] consists of 80% of document images and 20% of electronic documents. Another reason is that mobile devices with digital cameras are now coming into common use. Some users have already utilized such devices for taking digital copies of documents, because it is much more convenient than writing memo. This indicates that not only legacy documents but also new documents continue to be stored as document images.

In order to utilize such document images from the viewpoint of question answering, we have started a project of developing a QA system called "IQAS" (document Image Question Answering System). In this paper, we propose a method of *document image retrieval* for IQAS, by modifying our previous method [7], [8], which was developed as an extension of the method for electronic text [9], [10]. The major contribution of this paper is a way of improving *precision* of spotting parts that include the answer to the question. The previous method for document images, which is called the baseline method in this paper, employs *density distributions* of terms for retrieving appropriate parts of images. In this paper, new density distributions modified by taking into account the co-occurrence of successive terms in the question are introduced and tested by experiments on 1253 pages with 20 questions. The results of experiments show that the proposed method is superior to the baseline method.

The organization of this paper is as follows. In Sect. 2, we give an overview of the task of question answering for document images with some related work. Section 3 describes the proposed method of retrieving parts of document images with a mechanism of achieving higher precision. In Sect. 4, we experimentally evaluate the proposed method in comparison with the baseline method.

2. Question Answering for Document Images

2.1 Task and Configuration

The task of QA is *precision* oriented in nature. This is because the user is satisfied not by having all documents containing the same correct answer, but by just receiving the correct answer once. In the QA task, the user is allowed to ask questions in natural language. Systems for electronic text developed so far have tackled the questions of seeking simple *facts* by using "who", "what", "which", "when" and "where". Questions using "why" and "how" generally require much longer and complicated answers and thus their processing is still an open problem.

In order to locate facts in documents, QA systems are

Manuscript received October 9, 2004.
Manuscript revised April 11, 2005.

[†] The authors are with the Department of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, Sakai-shi, 599–8531 Japan.

[††] The author was with the Department of Computer and Systems Sciences, College of Engineering, Osaka Prefecture University, Sakai-shi, 599–8531 Japan.

[*] Presently, with the Graduate School of Information Science, Nara Institute of Science and Technology.

a) E-mail: kise@cs.osakafu-u.ac.jp

DOI: 10.1093/ietisy/e88–d.8.1843

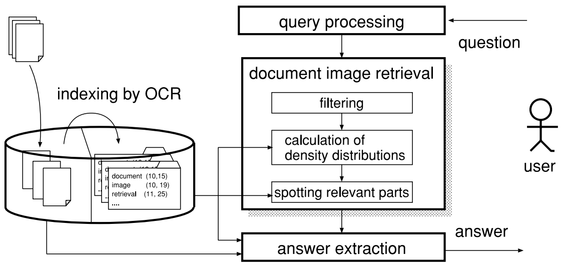

Fig. 1 System configuration.

generally based on the following configuration.

1. Query Processing : The question in natural language is analyzed to obtain both query terms and the type of question. Query terms are employed in the next step of processing. The type of question defines what the question asks about. For example, "location", "time" and "person" are typical types.
2. Document Retrieval : A document retrieval engine is employed to find documents which is likely to contain the answer to the question. Passage retrieval, i.e., to retrieve small portions of text from documents, is often utilized in this step.
3. Answer Extraction : The final step is to locate the answer in the retrieved passages with the help of types. Named entity extraction is applied to the extracted passages so as to locate the terms representing the answer to the question.

Our system IQAS also follows the above configuration. Figure 1 illustrates the system configuration of IQAS. In this paper we focus only on the second step, which is "document image retrieval" in our case.

2.2 Related Work

2.2.1 Passage Retrieval and QA for Electronic Text

Passage retrieval (e.g.,[10], [11]) and its application to QA (e.g.,[12], [13]) have been widely and actively pursued in the field of information retrieval. In passage retrieval for electronic text, it is common to take account of proximity of query terms: passages containing query terms densely would also include answers.

A problem encountered in finding such passages is that not all passages including answers contain *all* query terms [13]. A simple method to cope with this problem is to consider not *conjunction* but *disjunction* of query terms [13].

Our passage retrieval methods for electronic text [9], [10] and document images [7], [8] also have the above feature: the *density* of query terms is measured based on the summation of the density of each query term, which corresponds to taking the *disjunction* of query terms.

In this paper, we propose a new method of improving precision of document image retrieval by taking account of successive combinations of query terms as compared to the simple method based on the *summation*.

2.2.2 Document Image Retrieval

Document image retrieval has studied in both fields of information retrieval [14] and document image analysis [15]. One of the central issues has been how to cope with OCR errors. Other errors in higher level analyses such as layout analysis and logical labeling cause less influence to the retrieval results if retrieval systems are based on the well-known "bag of words" (BOW) model. This is because only the frequency of terms is utilized in the BOW model.

However, this does not hold for passage retrieval and question answering, since these are to segment parts defined based on the results of higher level analyses. Document images with recognized text in digital libraries often suffer errors of layout analysis: body text is mixed up with text in figures and tables as well as headers and footers. Moreover we sometimes encounter the case that textlines in different columns are failed to be separated. Thus methods for dealing with errors in higher level analyses are required.

Although a straightforward way is to improve the accuracy of high level analyses, we have taken an indirect way by proposing a different retrieval method [7], [8], which is an extension of our passage retrieval [9], [10] on electronic text to the two dimensional space. The characteristic point of the method is that it relies only on positions of terms in original pages. Parts are segmented not on the recognized text but on the two dimensional space of page regions. Density distributions of query terms, which were first introduced in [16], are employed for locating parts relevant to it. This enables us to retrieve parts of document images independently of the results of higher level analyses.

In this paper, we improve the above method of density distributions to be better suited for precision oriented retrieval.

3. Document Image Retrieval

3.1 Overview

The basic concept of the proposed method is to find parts of documents which densely contain terms in a query. The processing consists of the three steps shown in Fig. 1. Taking as input a set of index terms or *a query* extracted by the query processing, filtering is first applied to select pages which are likely to contain an answer to the question. Then the density distributions are calculated to find the parts which densely contain terms in the query. Finally, relevant parts are found based on the density distributions.

In the following, the details of each step are explained after a brief introduction of indexing and query processing.

3.2 Indexing

The process of indexing is basically the same as in our previous method. First, all words and their bounding boxes are extracted from page images with the help of OCR. Second, stemming and stopword elimination are applied to the extracted words[†]. The resultant words are called index terms (or simply terms) and stored with the centers of their bounding boxes. In other words, each page image is viewed as a two dimensional distribution of terms in it.

A new functionality introduced to the proposed method is the normalization of image size. In general, page images have various layouts. Some documents such as newspapers and technical journals may have multi-column layouts and thus densely contain a lot of terms in one page. On the other hand, others may have single-column layouts with a wider interline spacing and thus contain less terms. Documents with multi-column layouts would, therefore, be unevenly promoted if we simply computed the density of terms.

To avoid this harmful effect, it is necessary to normalize the size of page images. As the normalization constant C, we employ $C = H_m/5$ where H_m is the *mode* of textline height included in each document.

3.3 Query Processing

The task of query processing is both to identify the type of question as well as to extract index terms from the question. Suppose we have a query "Where is the Baseball Hall of Fame?". The query type is "location" from what it is asking and the index terms are "baseball", "hall" and "fame". Note that only the extraction of index terms is relevant to the task of document image retrieval. In the following, the sequence of extracted index terms is called the query and represented as $q = (q_1, \ldots, q_u)$ where q_i is called a query term and i indicates the order of occurrence in the question. For the above example, $q =$ (baseball, hall, fame).

3.4 Filtering

Filtering is applied to ease the burden of the next step which is relatively time-consuming. The task here is to select N_v pages that are likely to include the answer to the question.

For this purpose we utilize the simple vector space model (VSM) [17]. In the VSM, both a page p_j and a query q are represented as m-dimensional vectors:

$$\boldsymbol{p}_j = (w_{1j}, \ldots, w_{mj})^T \,, \tag{1}$$

$$\boldsymbol{q} = (w_{1q}, \ldots, w_{mq})^T \,, \tag{2}$$

where T indicates the transpose, w_{ij} is a weight of a term t_i in a page p_j, and w_{iq} is a weight of a term t_i in a query q. In this paper, we employ a standard scheme called "tf-idf" defined as follows:

$$w_{ij} = \mathrm{tf}_{ij} \cdot \mathrm{idf}_i \,, \tag{3}$$

where tf_{ij} is the weight calculated using the term frequency f_{ij} (the number of occurrences of a term t_i in a page p_j), and idf_i is the weight calculated using the inverse of the page frequency n_i (the number of pages containing a term t_i). In computing tf_{ij} and idf_i, the raw frequency is usually dampened by a function. We utilize $\mathrm{tf}_{ij} = \log(f_{ij} + 1)$ and $\mathrm{idf}_i = \log(n/n_i)$ where n is the total number of pages. The weight w_{iq} is similarly defined as $w_{iq} = \log(f_{iq} + 1)$ where f_{iq} is the frequency of a term t_i in a query q.

The similarity between a page p_j and a query q is measured by the cosine of the angle between $\boldsymbol{p}_j$ and $\boldsymbol{q}$:

$$\mathrm{sim}(\boldsymbol{p}_j, \boldsymbol{q}) = \frac{\boldsymbol{p}_j^T \boldsymbol{q}}{\|\boldsymbol{p}_j\| \, \|\boldsymbol{q}\|} \,. \tag{4}$$

where $\|\cdot\|$ is the Euclidean norm of a vector. Pages are sorted according to the similarity and top N_v pages are selected and delivered to the next step.

3.5 Calculation of Density Distributions

This step is to calculate density distributions for pages selected by the filtering. A density distribution of the query is defined based on that of each query term q_i. Let $T_i^{(p)}(x, y)$ be a weighted distribution of a term $q_i (= t_k)$ in a selected page p defined by:

$$T_i^{(p)}(x, y) = \begin{cases} \mathrm{idf}_k & \text{if } q_i (= t_k) \text{ occurs at } (x, y) \,, \\ 0 & \text{otherwise} \,, \end{cases} \tag{5}$$

where (x, y) is the center of the bounding box of a term. A density distribution $D_i^{(p)}(x, y)$ is a weighted distribution of q_i smoothed by a window $W(x, y)$:

$$D_i^{(p)}(x, y) = \sum_{u=-H_x}^{H_x} \sum_{v=-H_y}^{H_y} W(x - u, y - v) T_i^{(p)}(u, v) \,. \tag{6}$$

where $H_x = M_x/2$ and $H_y = M_y/2$. M_x and M_y are the horizontal and vertical widths of the window function, respectively. In the proposed method, we utilize a pyramidal window function shown in Fig. 2.

As discussed in 2.1, document image retrieval for the QA task should be precision oriented. An easy way of making retrieval precision oriented is to find parts which densely contain *all* query terms. This is achieved by the point-wise multiplication of corresponding density distributions:

$$C_u^{(p)}(x, y) = \prod_{i=1}^{u} D_i^{(p)}(x, y) \,, \tag{7}$$

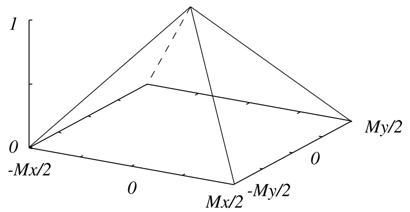

Fig. 2 Window function.

[†] Stemming is the process of normalizing words by keeping only *word stems*, e.g., from "processes" to "process". Stopwords are words that convey little meaning such as "a" and "the."

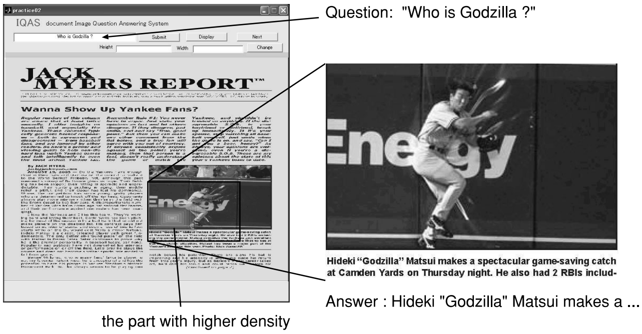
practice02
IQAS document Image Question Answering System
Who is Godzilla ?
Submit
Display
Next
Height
Width
Change
JACK MYERS REPORT™
Wanna Show Up Yankee Fans?
Question: "Who is Godzilla ?"
Hideki "Godzilla" Matsui makes a spectacular game-saving catch at Camden Yards on Thursday night. He also had 2 RBIs includ-
Answer : Hideki "Godzilla" Matsui makes a ...
the part with higher density

Fig. 3 Graphical user interface.

where u is the number of query terms.

However, this causes a problem in many cases because it is relatively rare that all query terms co-occur within a small region defined by the window function. In other words, $C_u^{(p)}(x, y)$ is zero if at least one of the density distributions $D_i^{(p)}(x, y)$ has the value of zero at (x, y).

A way to avoid this undesirable situation is to relax the requirement. In this paper, we consider the smaller number of successive query terms. For example, the density distribution obtained by $u - 1$ successive query terms is defined by

$$C_{u-1}^{(p)}(x, y) = \prod_{i=1}^{u-1} D_i^{(p)}(x, y) + \prod_{i=2}^{u} D_i^{(p)}(x, y) \,. \tag{8}$$

The reason for taking account of only the successive terms is that they are more relevant as compared to those randomly selected. For the general case of k successive query terms, the density distribution is defined by

$$C_k^{(p)}(x, y) = \sum_{j=0}^{u-k} \prod_{i=j+1}^{j+k} D_i^{(p)}(x, y) \,. \tag{9}$$

In the proposed method, the density distribution of the whole query for a page p is defined as the weighted sum of the combinations from all of the u terms down to s successive terms:

$$D^{(p)}(x, y) = \sum_{k=s}^{u} \alpha_k C_k^{(p)}(x, y) \,, \tag{10}$$

where the parameter s and the weight α_k are experimentally determined.

3.6 Spotting Relevant Parts

Based on the density distribution of Eq. (10), parts which are likely to include the answer are located on page images. First, page images are ranked according to their score of the maximum density:

$$s^{(p)} = \max_{x,y} D^{(p)}(x, y) \,. \tag{11}$$

Then, the top-ranked page is presented to the user through the GUI shown in Fig. 3. In this figure, the part with high density is highlighted. The user can magnify the retrieved part in the page. If it does not contain the answer, the user can retrieve the next page.

The proposed method of calculation of density distributions as well as spotting relevant parts may cause problems when relevant parts lie on multiple columns or pages. We simply ignore such cases since they are not frequent. In order to improve further the method, however, it is necessary to deal with such cases.

4. Experimental Results

4.1 Experimental Setting

The proposed method of document image retrieval was evaluated whether it enabled us to improve the precision as compared with the baseline method [7], [8]. A possible way of evaluation is end-to-end: their ability to find answers to queries is evaluated when they are employed in the QA system. In this scenario, it is necessary to utilize recognized documents including various kinds of errors such as OCR errors and errors of layout analysis. However such an end-to-end evaluation does not allow us to understand the contribution of each part of the QA system as well as the disturbance caused by each kind of errors in the data.

In this paper, we examine document image retrieval alone with restricted data as the first trial. The task here is to find passages which include answers to queries; no answer extraction process is applied to retrieved passages. As the data for evaluation we utilized terms with their bounding boxes that were easily obtained from PDF documents. We utilized only the information on terms because terms and their bounding boxes are often availabe even in the case that results of layout analysis are corrupted. In the experiments we utilized not the PDF documents with recognized text but PDF documents with original electronic text. Thus no OCR errors were included in the data. We consider that experiments with such clean data are appropriate to know the difference of the methods as the first trial; it is hopeless to improve the precision on recognized data if the proposed method gains no advantage on the clean data.

Needless to say it is inevitable to evaluate the proposed method with recognized data. For this purpose it is necessary to incorporate methods for coping with OCR errors [14], [15] into the proposed method. Moreover the evaluation in the end-to-end scenario is necessary to ensure that the proposed method is effective in the QA system. However these further evaluations are a part of our future work.

4.2 Data and Parameters

For the purpose of evaluating QA systems for electronic text, various test collections consisting of a large number of documents as well as questions are available (e.g., [2]). However it is impossible to utilize such data for our purpose because they are without the position of terms on pages.

Thus we decided to take PDF documents available on the Internet. Not only single column pages but also pages with multiple columns as shown in Fig. 3 were included. The topic of these PDF documents is the major league baseball. The number of documents and the total number of pages are 197 and 1253, respectively.

For the above documents, we prepared the queries shown in Table 1[†]. Answers to the queries are listed in Table 2. Some queries are associated with several possible answers delimited by commas; we regarded an output of the method as correct if at least one of them is included. Some answers consist of several terms like "setup man" for the query 11. In such cases, an output must include all of them to be regarded as correct. The parentheses in Table 2 indicate stopwords in the answers; these were not checked for marking.

Table 1 Queries used in the experiments.

Id	Query
1	What is the oldest stadium in Japan?
2	Who is Godzilla?
3	Who is the American League Leader in hits?
4	Who is the American League Leader in batting average?
5	Who is BRET BOONE?
6	From what are baseball gloves made?
7	From what are baseball bats made?
8	What variations are thrown in the major league?
9	Which team uses Koshien as home?
10	Who is Shigetoshi Hasegawa?
11	Where was Ichiro Suzuki born?
12	Where is the Baseball Hall of Fame?
13	Who is the world's best-known athletes?
14	Who is the most dominant and visible athlete in Japan?
15	Which stadium known as the House that Ruth Built?
16	What is First Aid Kit Rule?
17	What team does Mark McGuire play for?
18	What team did Babe Ruth play for?
19	What record is Mark McGwire close to breaking?
20	Which is the most famous stadium in Japan?

Table 2 Answers.

Id	Answer
1	Koshien
2	Matsui
3	Ichiro
4	Ichiro
5	(All-)star (second) baseman
6	cowhide
7	wood
8	seam fast ball, changeup, curveball, slider, split finger, forkball, knuckleball
9	Hanshin
10	setup man
11	Japan, Tokyo, Honshu
12	New York
13	Sosa, Jeter, Piazza, Rordriguez
14	Ichiro
15	Yankees
16	first aid kit
17	Cardinals
18	New York Yankees
19	(the most) homeruns (in one) season
20	Koshien

The above pairs of queries and answers were prepared by a student who read through all PDF documents. Thus it is guaranteed that at least one answer is included in the PDF documents.

Table 3 lists the ranges of parameters tested in the experiments. As the unit of length for the window size, 1/5 of the mode of textline height in each document is utilized. In the experiments, the window size varied from the height of 3.6 (=18/5) textlines to 20 (=100/5) textlines. The value of s indicates the minimum number of combined successive terms. Thus if a query includes five terms $(q_1, \dots, q_5)$ and $s = 2$ is applied, the successive combinations of all terms $(q_1, \dots, q_5)$ down to two terms (q_1, q_2), (q_2, q_3), $(q_3, q_4), (q_4, q_5)$ are considered in calculating the density distributions. As for the value of α_k, we tested "1" (equal weight) and "k" (varied weight). Since k corresponds to the number of combined terms, combinations with a larger number of terms are more important in the case of $\alpha_k = k$. The number of pages N_v selected at the filtering was fixed to 10 throughout the experiments.

4.3 Evaluation

4.3.1 Reciprocal Rank

The output of the method is the ranked pages with their density distributions. We regarded a page as correct in case a correct answer listed in Table 1 is found in the N_t nearest terms to the peak of the density distribution in the page. For each query, top *five* pages obtained by the method are verified whether they are correct.

The results were evaluated using the score "reciprocal ranks". The reciprocal rank for a query is calculated as $1/r$ where r is the rank of the page which first contains the correct answer. For example, if the third-ranked page first contains the correct answer, the reciprocal rank is 1/3. We also utilize the score "mean reciprocal rank" (MRR) defined as the average of reciprocal ranks for all queries.

4.3.2 Leave-One-Out Cross Validation

Experiments were carried out by applying the *leave-one-out cross validation*, i.e., values of parameters were selected by training based on the all but one *left-out* query and the selected values were applied to the processing of the left-out query as a test. By altering the left-out query, we obtain reciprocal ranks whose number is the same as that of the queries. MRR is then obtained by averaging the resultant reciprocal ranks.

4.3.3 Baseline Method

For the purpose of comparison, we applied the simplest vari-

[†] The queries 13, 14 and 20 are to seek answers beyond the category called "factoid." Although not all users agree to the answers listed in Table 2, we considered them as correct simply because there existed such descriptions (or opinions) in the documents for experiments.

Table 3 Parameters and their ranges.

Parameter		Range
width of the window	M_x	18 ∼ 100 step 4 †
height of the window	M_y	18 ∼ 100 step 4 †
min. no. of terms combined in Eq. (10)	s	1 ∼ total no. of terms in a query
the weight for $C_k^{(p)}(x, y)$	α_k	$\equiv 1$ or $= k$

† measured in units of 1/5 of the mode of textline height.

Table 4 MRR and values of parameters obtained by training.

	N_t	MRR	M_x		M_y		s	α_k
			ave.	mode	ave.	mode		
proposed	30	0.626	67.6	70	75.8	78	2	k
method	10	0.579	58	58	77.6	78	2	k
baseline	30	0.503	42.4	38	38	38	—	—
method	10	0.490	33.8	30	33.4	30	—	—

ant of our previous method [7], [8] as the *baseline*. In this method, density distributions are calculated based not on Eq. (10) but on the following:

$$D^{(p)}(x, y) = \sum_{i=1}^{u} D_i^{(p)}(x, y) . \tag{12}$$

Except for this difference, all processing steps are shared with the proposed method.

4.4 Statistical Test

The evaluation here is to compare values of MRR obtained by the two methods. An important question is whether the difference in MRR is really meaningful or just by chance. In order to make such a distinction, it is necessary to apply a statistical test.

Several statistical tests have been applied to the task of information retrieval [18], [19]. In this paper, we utilize a test called "the paired t-test" [18] (called " macro t-test" in [19]) The following is the summary of the test.

Let a_i and b_i be scores (e.g., reciprocal ranks) by methods A and B for the same query i , and define $d_i = a_i - b_i$. The test can be applied under the assumptions that the model is additive, i.e., $d_i = \mu + \varepsilon_i$ where μ is the population mean and ε_i is an error, and that the errors are normally distributed. The null hypothesis here is $\mu = 0$ (A performs equivalently to B in terms of the scores), and the alternative hypothesis is $\mu > 0$ (A performs better than B).

It is known that the Student's t-statistic

$$t = \frac{\bar{d}}{\sqrt{s^2/N}} \tag{13}$$

follows the t-distribution with the degree of freedom of $N-1$, where N is the number of samples (queries), $\bar{d}$ and s^2 are the sample mean and variance:

$$\bar{d} = \frac{1}{N} \sum_{i=1}^{N} d_i , \tag{14}$$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (d_i - \bar{d})^2 . \tag{15}$$

By looking up the value of t in the t-distribution, we can obtain a p-value, i.e., the probability of observing the sample results d_i ($1 \leq i \leq N$) under the assumption that the null hypothesis is true. The p-value is compared to a predetermined significance level in order to decide whether the null hypothesis should be rejected. As significance levels, we utilize 0.05 and 0.01.

4.5 Results and Discussion

4.5.1 Training

Let us first show the results of training. Table 4 shows MRR obtained through the training. As the number of nearest terms N_t, which is related to the accuracy of results, we utilized 30 and 10; the task is harder with a smaller N_t. As shown in this table, the proposed method outperformed the baseline method for both values of N_t.

Values of parameters selected at the training are also listed in Table 4. Let us next discuss the values of s and α_k, both of which are only for the proposed method. The proposed method often performed best with $s = 2$ and $\alpha_k = k$ for both N_t's. The value $s = 2$ indicates that it is better not to take into account the distributions of single terms. As stated in Sect. 3.5, s indicates the requirement of "co-occurrence" of successive terms within the window region. The baseline method is, on the other hand, to calculate the density distributions by taking into account only the single terms (see Eq. (12)). Thus the results indicate that the co-occurrence plays an important role for locating the answer accurately. The selection of $\alpha_k = k$ means that the "co-occurrence" with a larger number of terms is more important than those with less terms.

Let us turn to the window widths. Table 4 shows that (1) smaller windows are required for smaller N_t for locating the answers more accurately, and (2) the baseline method requires smaller windows as compared to the proposed method. Because smaller windows provide us less capability of smoothing the distributions, they are not desirable from the viewpoint of stability of the processing. For example, the baseline method with $N_t = 10$ uses the window

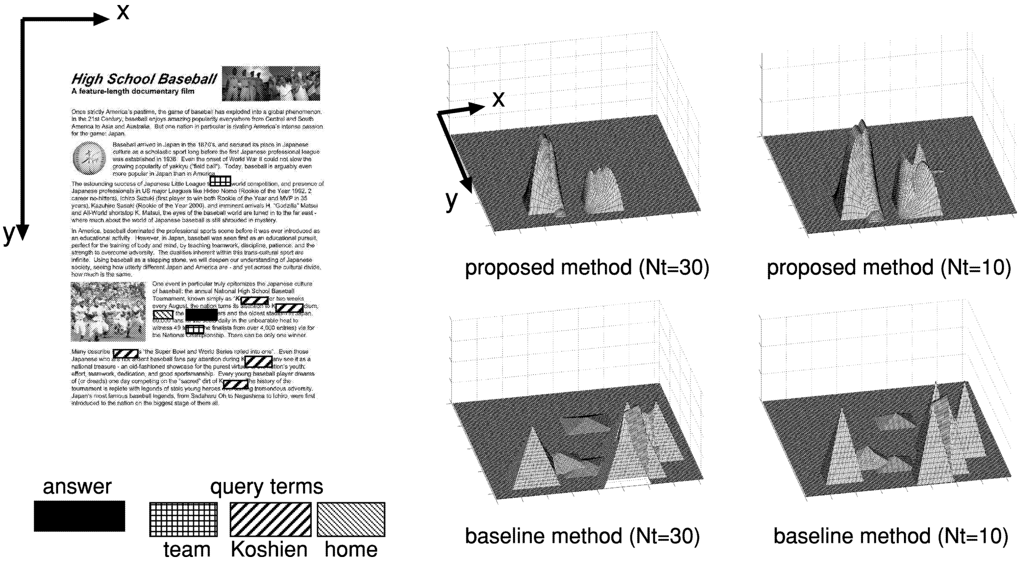

Fig. 4 Examples of density distributions.

Table 5 Results of test.

	N_t	query Id																				MRR
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
proposed method	30	1	1	1	1	1	1	1	0	$\frac{1}{2}$	1	1	0	0	1	0	1	0	0	0	0	0.575
	10	$\frac{1}{2}$	1	1	$\frac{1}{2}$	1	1	0	0	0	1	1	0	0	1	0	1	0	0	0	0	0.450
baseline method	30	$\frac{1}{3}$	1	1	$\frac{1}{3}$	0	$\frac{1}{4}$	1	0	0	0	$\frac{1}{5}$	0	0	1	0	$\frac{1}{2}$	0	0	0	$\frac{1}{3}$	0.298
	10	$\frac{1}{2}$	1	1	1	0	$\frac{1}{4}$	1	0	0	1	0	0	0	1	0	$\frac{1}{2}$	0	0	0	$\frac{1}{3}$	0.379

of size 30×30. Since the typical height of body textlines are normalized to 5, the window is of size 6 lines. On the other hand the proposed method employs windows of size 12 to 16 textlines.

Examples of density distributions are illustrated in Fig. 4. The baseline method yielded some spikes in the distributions. On the other hand, the proposed method generated smooth distributions. In general, larger windows allow us to obtain smoothness, though they spoil the accuracy of locating the answers. The proposed method avoids this side effect by using the combinations of terms.

4.5.2 Test

Table 5 shows the results of test for the left-out queries. In this table, "query Id" indicates the left-out query and the numbers for them represent the reciprocal ranks.

For the queries 8, 12, 13, 15, and 17–19, neither of the methods could find the answers within top five pages. This was partly due to repetitious use of general query terms in pages. For example, the query 8 includes the terms "variation", "throw", "major" and "league" all of which are commonly used in documents on the major league baseball. Another and more important reason is that the methods are without the "filtering" capability based on the type of queries. For instance, the query 12 asks the location but only one page among all top five pages (in total 20 pages) included the name of location. Filtering would allow us to solve the problem as in the systems for electronic text.

For the queries 2, 3 and 14, both of the methods found the answers in the top ranked page. The difference between the methods was caused by the rest.

For the query 20, the proposed method was inferior to the baseline method. This was caused by the erroneous selection of the value of s in the proposed method. In this case, $s = 3$ is selected by the training. Since the query includes three terms, the selection indicates the requirement of co-occurrence of all terms. But unfortunately, no page included all within the size of the window.

For the queries 5, 6, 9, 11 and 16, on the other hand, the proposed method outperformed the baseline method. In most of these cases, the baseline method yielded erroneous results due to the repetitious use of some terms in the query. For example, a page including the term "glove" frequently was erroneously ranked at the top by the baseline method, though the query also includes the terms "baseball" and "made". For these cases, therefore, the proposed method which put additional weights for co-occurrence of terms was successful.

In total, the values of MRR show that the proposed method outperformed the baseline method for both values of N_t.

Now the question is whether the above advantages of the proposed method are statistically significant. The results

Table 6 Results of the statistical test.

method A	N_t	method B	N_t	results
proposed method	30	baseline method	30	$\gg$
			10	$>$
	10		30	$\sim$
			10	$\sim$
proposed method	30	proposed method	10	$>$
baseline method	30	baseline method	10	$\sim$

$\gg$: p-value $\leq$ 0.01
$>$: 0.01 $<$ p-value $\leq$ 0.05
$\sim$: 0.05 $<$ p-value

of the statistical test are shown in Table 6. The meaning of the symbols such as "≫", ">" and "∼" is summarized at the bottom of the table. For example, the symbol "≫" was obtained in the case of the proposed method with $N_t = 30$ compared to the baseline method with the same N_t. This indicates that, at the significance level 0.01, the null hypothesis "the proposed method with $N_t = 30$ performs equivalently to the baseline method with $N_t = 30$" is rejected and the alternative hypothesis "the proposed method performs better than the baseline method" is accepted. Roughly speaking, "$A \gg B$", "$A > B$" and "$A \sim B$" indicate that "A is almost guaranteed to be better than B", "A is likely to be better than B" and "A is equivalent to B", respectively.

As shown in Table 6, the proposed method with $N_t = 30$ is always significantly better than other methods. This indicates the advantage of the proposed method. In the cases with $N_t = 10$, however, the proposed method is not significantly better than the baseline method. Thus additional experiments are required to clarify whether or not the proposed method also outperforms the baseline method with smaller N_t's. In order to obtain more concrete evidence of the advantage, it is required to employ a larger number of queries and documents.

5. Conclusion

In this paper we have presented a method of document image retrieval that is precision oriented for the task of question answering. The characteristic point of the method is that it takes combinations of successive query terms into account when calculating density distributions. This allows us to improve the accuracy of locating answers without increasing the spurious spikes in the distributions. From the experimental results we confirmed that the proposed method with $N_t = 30$ outperformed the baseline method.

Future work includes experiments with a larger number of queries and documents, as well as with OCR'ed documents. The implementation of the whole system with the capabilities of "query type identification" and "answer extraction" is also important future work.

Acknowledgement

This research was supported in part by Grant-in-Aid for Scientific Research (C) from Japan Society for the Promotion of Science (No.14580453) and Casio Science Promotion Foundation.

References

[1] E.M. Voorhees, "Overview of the TREC 2002 question answering track," Proc. Text REtrieval Conference 2002, http://trec.nist.gov/pubs/trec11/t11_proceedings.html

[2] http://trec.nist.gov/

[3] C.C.T. Kwok, O. Etzioni, and D.S. Weld, "Scaling question answering to the Web," Proc. WWW10, pp.150–161, 2001.

[4] http://www.ai.mit.edu/projects/infolab/

[5] http://labs.nttrd.com/ (in Japanese)

[6] http://www.acm.org/dl/

[7] K. Kise, M. Tsujino, and K. Matsumoto, "Spotting where to read on pages — Retrieval of relevant parts from page images," Proc. DAS'02, pp.388–399, 2002.

[8] K. Kise, W. Yin, and K. Matsumoto, "Document image retrieval based on 2D density distributions of terms with pseudo relevance feedback," Proc. ICDAR 2003, pp.488–492, 2003.

[9] H. Mizuno, K. Kise, and K. Matsumoto, "Linking figures and tables to their expository texts using word density distributions and their biases," J. IPSJ, vol.40, no.12, pp.4400–4404, 1999.

[10] K. Kise, M. Junker, A. Dengel, and K. Matsumoto, "Passage retrieval based on density distributions of terms and its applications to document retrieval and question ansewering," in Reading and Learning, LNCS 2956, eds. A. Dengel, M. Junker, and A. Weisbecker, pp.306–327, 2004.

[11] O. de Kretser and A. Moffat, "Effective document presentation with a locality-based similarity heuristic," Proc. SIGIR'99, pp.113–120, 1999.

[12] C.L.A. Clarke and E.L. Terra, "Passage retrieval vs. document retrieval for factoid question answering," Proc. SIGIR'03, pp.427–428, 2003.

[13] H. Isozaki, "NTT's question answering system for NTCIR QAC2," Working Notes of NTCIR-4, pp.326–332, 2004.

[14] J. Callan, P. Kantor, and D. Grossman, "Information retrieval and OCR: From converting content to grasping meaning," SIGIR Forum, vol.36, no.2, pp.58–61, 2002.

[15] D. Doermann, "The indexing and retrieval of document images: A survey," Comput. Vis. Image Process., vol.70, no.3, pp.287–298, 1998.

[16] S. Kurohashi, N. Shiraki, and M. Nagao, A method for detecting important descriptions of a word based on its density distribution in text," Trans. IPSJ, vol.38, no.4, pp.845–853, 1997.

[17] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, 1999.

[18] D. Hull, "Using statistical testing in the evaluation of retrieval experiments," Proc. SIGIR'93, pp.329–338, 1993.

[19] Y. Yang and X. Liu, "A re-examination of text categorization methods," Proc. SIGIR'99, pp.42–49, 1999.

Koichi Kise received the B.E., M.E., and Ph.D. degrees in communication engineering from Osaka University, Osaka, Japan, in 1986, 1988 and 1991, respectively. From 2000 to 2001, he was a visiting professor at German Research Center for Artificial Intelligence (DFKI), Germany. He is now a Professor of the Department of Computer Science and Intelligent Systems, Osaka Prefecture University, Japan. His research interests include document image analysis, information retrieval, and image processing.

Shota Fukushima received the B.E. degree in computer and systems sciences from Osaka Prefecture University in 2004. He is currently a graduate student of the Nara Institute of Science and Technology. He worked on document image retrieval for his graduation thesis.

Keinosuke Matsumoto is a Professor of the Department of Computer Science and Intelligent Systems, Osaka Prefecture University. Previously, he worked for Mitsubishi Electric Corporation as a researcher. He received the Ph.D. degree in electrical engineering from Kyoto University for his work on a knowledge-based approach to power system restoration. His research interests include software engineering, object-oriented technology, and intelligent systems.