IJE TRANSACTIONS A: Basics Vol. 27, No. 1, (January 2014) 7-14

International Journal of Engineering

Journal Homepage: www.ije.ir

Document Image Retrieval Based on Keyword Spotting Using Relevance Feedback

M. Keyvanpour* [a], R. Tavoli[b], S. Mozaffari[c]

[a] *Department of Computer Engineering, Alzahra University, Tehran, Iran*
[b] *Department of Electrical and Computer, Islamic Azad University, Qazvin Branch, Qazvin, Iran*
[c] *Electrical and Computer Engineering Departmet, Semnan University, Semnan, Iran*

PAPER INFO

Paper history:
Received 15 March 2013
Received in revised form 11 June 2013
Accepted 20 June 2013

Keywords:
Relevance Feedback
Document Image
Information Retrieval
Keyword Spotting

ABSTRACT

Keyword Spotting is a well-known method in document image retrieval which is based on query word image. In this paper, a document image retrieval system based on keyword spotting and relevance feedback is presented. Relevance feedback as an interactive method is used in this paper to improve the performance of Document Image Retrieval System (DIRS). In the proposed method, we compare several strategies of positive and negative feedbacks which include "Only Positive Feedback", "Only Negative Feedback" and "Positive and Negative Feedback". Experiments show that using relevance feedback in DIR outperforms common DIR.

doi: *10.5829/idosi.ije.2014.27.01a.02*

1. INTRODUCTION

Document Image Retrieval System (DIRS) based on keyword spotting performs matching process directly on data images using word-images as queries. It usually compares common features, such as width to height ratio, word area density and shape projections, all extracted from the word document image. In recent years, several attempts have been made by researchers to retrieve document images using word images. A detailed survey on document image retrieval up to 1997 can be found in research article [1]. An overview on document image retrieval system is presented by Kokare and Shirdhonkar [2]. Keyvanpour and Tavoli have proposed a framework for classification and evaluation of document image retrieval approachs [3]. In this framework, the methods are classified into two groups: text components methods and non-text components. Word level image matching and retrieval for printed documents are presented in some references [4-11].

In previous works on word shape coding, Li et al. [5] used an alternative technique and combination of feature descriptors for keyword spotting without the use of OCR. Lu and Tan [6] proposed a system for designing an information retrieval system with ability of dealing with image document stored in digital libraries. A novel partial matching algorithm is designed by Meshesha and Jawahar [7] for morphological matching of word form variants in a language. Leydier et al. [8] used DIP techniques to create a pattern dictionary of each document and performed word spotting by selecting gradient angle feature and a matching algorithm. Lu et al. [9] annotated word images using a set of topological shape features including character ascenders/descenders, character holes and character water reservoirs. With the annotated word shape codes, document images can be retrieved by either query keywords or query image. A set of document image processing techniques extracting powerful features for word image description has been presented in [10]. Keyvanpour and Tavoli [11] peoposed a feature weighting method to improve DIRS performance. In this method, they weighted feature using coefficient of multiple correlations.

A key requirement for developing future document image retrieval systems is to explore the synergy between humans and computers. Relevance feedback (RF) is a technique that engages the user and the

*Corresponding Author Email: keyvanpour@alzahra.ac.ir (M. Keyvanpour)

retrieval system in a process of symbiosis [12]. The idea of RF usage in information retrieval systems is to adapt the system to the specific user preferences making more important weights or features that reflect the actual user needs in order to achieve higher precision. Therefore, we can define relevance feedback as the process by which human and computer interact in order to automatically adjust an existing query to the real user preferences. Research has been devoted in the past few years to relevance feedback as an effective solution to improve performance of information retrieval system [12-18]. A comprehensive review on RF in image retrieval is presented [13]. MacArthur et al. presented a relevance feedback technique that uses decision trees to learn a common thread among instances marked relevant [12]. Rota Bulò et al proposed a novel approach to content-based image retrieval with relevance feedback, which is based on the random walker algorithm introduced in the context of interactive image segmentation [14]. The idea is to treat the relevant and non-relevant images labeled by the user at every feedback round as "seed" nodes for the random walker problem. Su et al. [15] proposed a new feedback approach with progressive learning capability combined with a novel method for feature subspace extraction. The proposed approach is based on a Bayesian classifier and utilizes positive and negative feedbacks examples with different strategies.

In this paper, we propose the use of RF method to improve DIRS accuracy. First, architecture of the proposed system is presented. Then, each building block is described in more details. In this paper, we compare a variety of strategies for positive and negative feedbacks which include "Positive Feedback", "Negative Feedback" and "Positive and Negative Feedback". We evaluate the proposed system using precision and recall measures. This paper is organized as follows. Section 2 describes document image retrieval and RF concepts. Section 3 presents the proposed system. Section 4 explains evaluation measures used in this paper. Experimental results of the proposed system are presented in section 5. Finally, concluding remarks are given in section 6.

2. DOCUMENT IMAGE RETRIEVAL AND RELEVANCE FEEDBACK CONCEPTS

Since the proposed document image retrieval (DIR) system is based on relevance feedback (RF), these concepts are briefly discussed in this section.

2. 1. Document Image Retrieval Figure 1 depicts the overall structure of the DIR system base on word spotting [4]. It is composed of two main parts: the offline and the online operations. In the offline operation, the archive of document images are examined and the results are stored in a database. This digital "scanning" consists of three stages. At first stage, the document passes the preprocessing stage which includes a binarization with the Otsu method, a mean filter and a skeletonization operation.

After preprocessing, word segmentation stage is performed. Its primary goal is to detect the word blocks. At the final stage of the offline operation, features of each word are calculated and stored in the database [4]. For each word block, a total of 7 different features are extracted: width to height ratio, word area density, center of gravity, vertical projection, top–bottom shape projections, upper grid features and down grid features.

The online operation consists of the interface from which the user can manipulate the system (enter the query word and see the results), creation of the word's image, preprocessing and feature extraction stages which are the same with that in the offline operation and finally, the matching process of the query word's features with them in the database.

2. 2. Relevance Feedback Relevance feedback, originally developed for information retrieval [16], is a supervised learning technique used to improve the effectiveness of information retrieval systems. The main idea of RF is using positive and negative samples provided by the user to enhance the system's accuracy. Positive and negative samples are retrieved relevant and non-relevant documents by user, respectively. For a given query, the system first retrieves a list of ranked images according to the predefined similarity metrics, which are often defined as the distance between feature vectors of images.

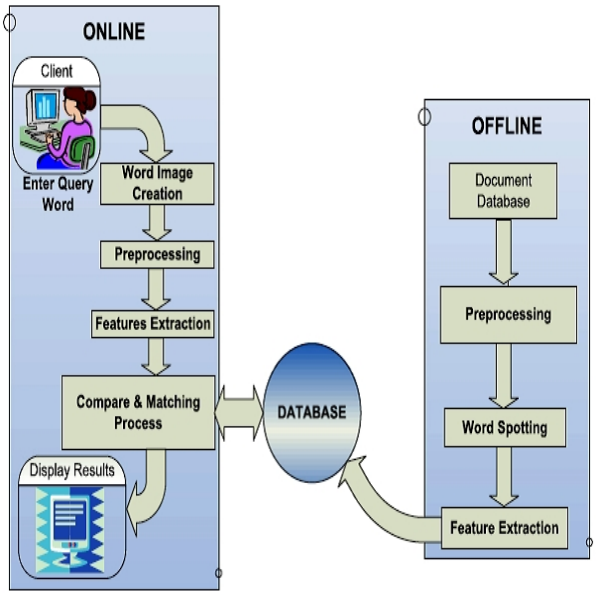

Figure 1. Overall structure of document image retrieval system [4].

Then, the user selects a set of positive and/or negative examples from this list. Then, system refines the query and retrieves a new list of images. The original relevance feedback method, in which the vector model is used for document retrieval, can be illustrated by the Rocchio's formula [16] as:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{w}_j \in D_r} \vec{D}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{w}_j \in D_{nr}} \vec{D}_j \quad (1)$$

where q_m is the modified query, q_0 is the original query vector, D_r and D_{nr} are the set of known relevant and non-relevant documents, respectively and α, β, and γ are weight terms.

3. PROPOSED SYSTEM

In this paper, we propose the use of relevance feedback method to improve DIRS accuracy. System architecture is shown in Figure 2.

In the proposed method, first, the user enters a word image query. Then, the query feature vector is created. For each word block, a total of 7 different features are extracted: width to height ratio, word area density, center of gravity, vertical projection, top–bottom shape projections, upper grid features and down grid features. Then, a descriptor is created by the seven extracted features as shown in Table 1. The first element is the weight to height feature; the second one is the image area density feature and the third one is the center of gravity feature. The following twenty elements are the ones extracted from the vertical projection feature and the next fifty from the top–bottom shape projection features. Finally, the last twenty elements are the ones extracted from the upper and down grid features divided by 10 in order to prevent overpowering the other features. The rest of the features values are normalized from 0 to 1.

After that, the query feature vector is compared with indexed words in the database. Minkowski distance between query feature vector and indexed words is used for this purpose [4]:

$$MD(i) = \sum_{k=1}^{93} |Q(k) - W(k,i)| \quad (2)$$

$$R_i = 100\left(1 - \frac{MD(i)}{\max(MD)}\right) \quad (3)$$

where $MD(i)$ is the Minkowski distance of the i word, $Q(k)$ is the query descriptor and $W(k, i)$ is the descriptor of the ith word. Then, the similarity rate of the remaining words is computed. The rate is a normalized value between 0 and 100, which depicts how similar the words of the database are with the query word. R_i is the rate value of the word i, and $max\,(MD)$ is the maximum Minkowski distance found in the document database.

TABLE 1. The structure of descriptor

Position	Features
1st Position	Width to height
2nd Position	Image area
3rd Position	Center of gravity
4th-23rd Position	Vertical projection
24th-48th Position	Top shape projection
49th-73rd Position	Bottom shape projection
74th-83rd Position	Upper grid
84th -93rd Position	Down grid

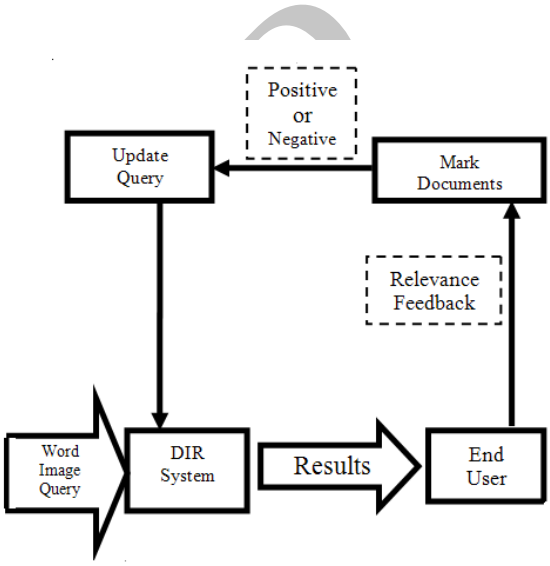

Figure 2. Proposed system

The system presents retrieval results according to the distance measured. Then, the user selects a set of positive and/or negative examples from the retrieved document images. Subsequently, the system refines the query and retrieves a new list of documents. This paper compares variety of strategies for positive and negative feedback which include "Only Positive Feedback", "Only Negative Feedback" and "Positive and Negative Feedback". For the selected positive feedback, as relevant to all the word images from the initial query, the relevent results would be judged by user. For positive feedback Rocchio's formula is changed to:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|W_r|} \sum_{\vec{W}_j \in W_r} \vec{W}_j \quad (4)$$

For negative feedback, non-relevant word images from the initial query result have been selected according to the user judgment. For negative feedback, Rocchio's formula is also modified:

$$\vec{q}_m = \alpha \vec{q}_0 - \gamma \frac{1}{|W_{nr}|} \sum_{\vec{W}_j \in W_{nr}} \vec{W}_j \quad (5)$$

The same process is performed for positive and negative feedback:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|W_r|} \sum_{\vec{w}_j \in W_r} \vec{W}_j - \gamma \frac{1}{|W_{nr}|} \sum_{\vec{w}_j \in W_{nr}} \vec{W}_j \quad (6)$$

In Equations (4), (5) and (6), q_0 is the original query vector, W_r and W_{nr} are the set of known relevant and non-relevant words in documents, respectively and α, β, and γ are weight terms, respectively.

4. EVALUATION MEASURES

Precision, recall and F-measure are widely used for evaluation of the document image retrieval system [4]. Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. In our evaluation, the precision and recall values are expressed in percentage. They are defined in Equations (7) and (8) as follows:

$$precision = \frac{\#(\text{Re}\ levant\ \text{Re}\ trieved\ \text{Re}\ cords)}{TotalNumbe\ rOf\ \text{Re}\ trieved\ \text{Re}\ cords} \quad (7)$$

$$\text{Re}\ call = \frac{\#(\text{Re}\ levant\ \text{Re}\ trieved\ \text{Re}\ cords)}{TotalNumbe\ rOf\ \text{Re}\ levant\ \text{Re}\ cords} \quad (8)$$

A single measure that trades off precision versus recall is the F-measure which is the weighted harmonic mean of precision and recall:

$$F - Measure = \frac{2.(precision \times recall)}{(precision + recall)} \quad (9)$$

In our evaluation the precision and recall and F-measure values are expressed in percentage.

5. EXPERIMENTAL RESULTS

In our experiments, the evaluation of the proposed system was based on 100 document images. The database of the documents has been created automatically from various digital text documents. In order to calculate the precision and recall, 30 searches were made using random words. In this paper, we tested the system with several strategies for positive and negative feedback. For obtaining the best amount of α, β in positive feedback, we tested difference values of α, β to earn optimum values of precision and recall. In positive feedback, we set Rocchio's formula with α=1 and β=0.82. In positive feedback, precision and recall values obtained are depicted in Figure 4 (a) and (b), respectively.

As shown in Figure 4, using positive feedback, performance of DIRS in term of average precision is increased while average of recall is fixed. Positive feedback outweigh the negative feedback for the two following reasons. Positive feedback converges modified query to the relevant documents. But negative feedback may not converge to the relavent documents. So most information retrieval systems set $\gamma<\beta$. β, γ are weights of positive and negative feedback, respectively. Experiments also show that use of positive feedback in DIR achieves better performance than common DIR with no feedback. Table 1 compares the average precision and recall of the proposed approach with DIRS [4] and WDIRS [10].

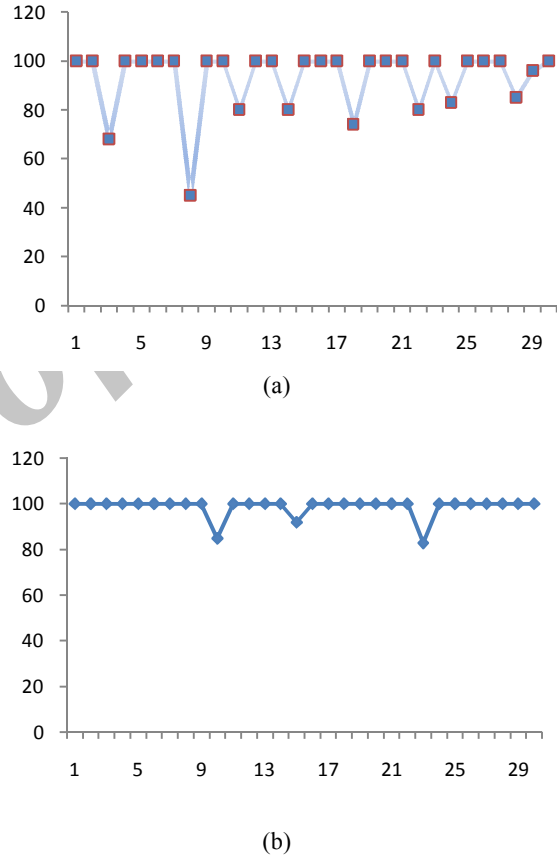

Figure 3. (a) The variation of the precision coefficient of the proposed method (Positive Feedback) for 30 searches.The average precision is 93.03%. (b) The variation of the recall coefficient of the proposed method (Positive Feedback) for 30 searches.The average precision is 98.66%.

TABLE 2. Comparison of the average precision and recall between proposed system and DIRS and WDIRS

	Precision	Recall	F-measure
DIRS [4]	87.8%	99.26%	93.03%
WDIRS [10]	55.43%	94.78%	69.95%
Positive feedback in DIRS	93.03%	98.66%	95.76%

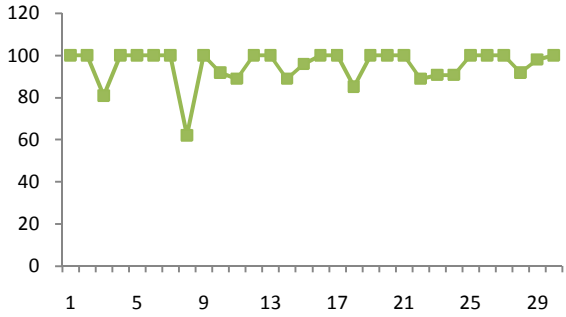

Figure 4. The variation of the F-measure coefficient of the proposed method (Positive Feedback) for 30 searches. The average F-measure is 95.76%.

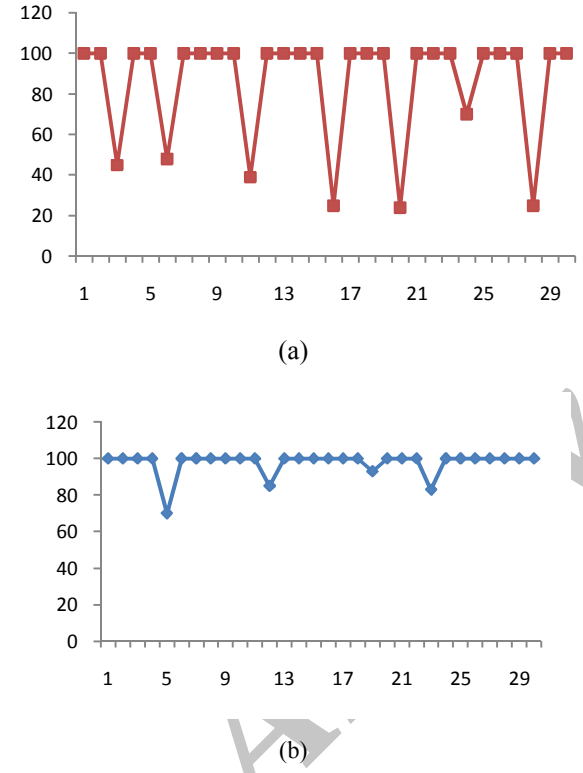

Figure 5. (a) The variation of the precision coefficient of Negative Feedback for 30 searches. The average precision is 85.86%. (b) The variation of the recall coefficient of the Negative Feedback for 30 searches. The average precision is 97.7%.

As shown in Table 2, the average precision in WDIRS and DIRS is 55.43% and 87.8%, respectively. Also, average recall in WDIRS and DIRS is 94.78% and 99.26%, respectively. After using positive feedback in DIRS, the average precision is 93.03% and the average recall becomes 98.66%. With selecting positive samples by user, the results of modified query would be closer to the relevant documents. Then, total number of retrieved documents decreases and according to Equation (7) precision increases. Also, with decreasing the number of retrieved documents, the number of retrieved relevant documents less decreased. Finally according to Equation (8) the amount of recall less decreased. Comparing proposed method with WDIRS [10], both precision and recall have increased in our system. But comparing with DIRS [4], the proposed method has more precision but less recall. In spite of degradation of recall, F-measure criteria enhanced that shows an improvement by the proposed method.

According to Figure 6, using negative feedback in DIRS, performance of DIRS in terms of average precision and recall decreased. Because negative relevance feedback is a special case where we do not have any positive example; this often happens when the search results are poor. So, negative feedback is not suitable and has a less efficiency than DIRS [4].

In Figure 7, by using both positive and negative feedbacks in DIRS, performance of DIRS in term of average precision is increased while average of recall is decreased. However, the average of F-measure has increased.

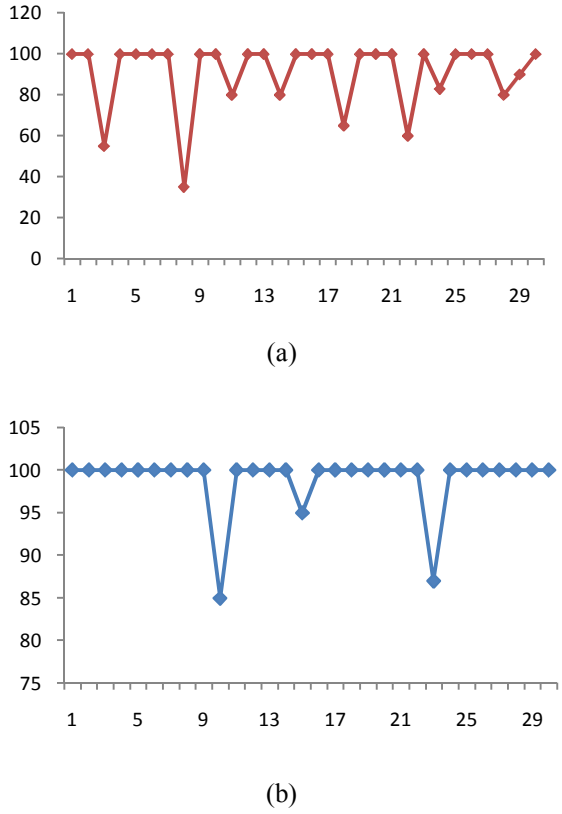

Figure 6. (a) The variation of the precision coefficient of Positive and Negative Feedback for 30 searches. The average precision is 90.93 %. (b) The variation of the recall coefficient of the Positive and Negative Feedback for 30 searches. The average precision is 98.9%.

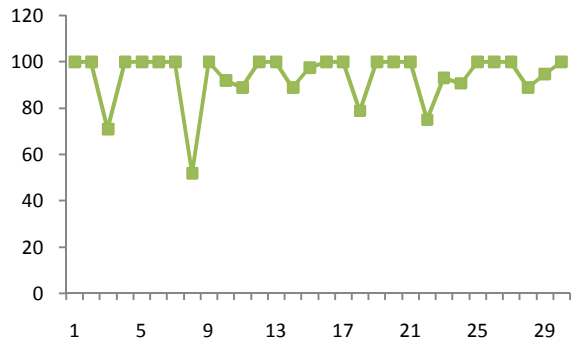

Figure 7. The variation of the F-measure coefficient of the proposed method (Positive and Negative Feedback) for 30 searches. The average of F-measure is 94.7%.

6. CONCLUSION

In many information retrieval systems, relevance feedback is used to increase accuracy. In this paper, we use this technique to improve document image retrieval system performance. This paper compares a variety of strategies for positive and negative feedbacks. These are "Only Positive Feedback", "Only Negative Feedback" and "Positive and Negative Feedback". Experiment results show that using RF, especially positive feedback, in DIR outperforms common DIR.

7. REFERENCES

1. Doermann, D., "The indexing and retrieval of document images: A survey", ***Computer Vision and Image Understanding***, Vol. 70, No. 3, (1998), 287-298.
2. Kokare, M. B. and Shirdhonkar, M., "Document image retrieval: An overview", ***International Journal of Computer Applications***, Vol. 1, No. 7, (2010), 114-119.
3. Keyvanpour, M. and Tavoli, R., "Document image retrieval: Algorithms, analysis and promising directions", ***International Journal of Software Engineering and Its Applications***, Vol. 7, No. 1, (2013), 93-106.
4. Zagoris, K., Ergina, K. and Papamarkos, N., "A document image retrieval system", ***Engineering Applications of Artificial Intelligence***, Vol. 23, No. 6, (2010), 872-879.
5. Li, L. Bai, S.,and Tan, C. L., "Keyword spotting in document images through word shape coding", in Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on, IEEE. (2009), 331-335.
6. Lu, Y. and Tan, C. L., "Information retrieval in document image databases", ***Knowledge and Data Engineering, IEEE Transactions on***, Vol. 16, No. 11, (2004), 1398-1410.
7. Meshesha, M. and Jawahar, C., "Matching word images for content-based retrieval from printed document images", ***International Journal of Document Analysis and Recognition (IJDAR)***, Vol. 11, No. 1, (2008), 29-38.
8. Leydier, Y., LeBourgeois, F. and Emptoz, H., "Textual indexation of ancient documents", in Proceedings of the 2005 ACM symposium on Document engineering, ACM. (2005), 111-117.
9. Lu, S., Li, L. and Tan, C. L., "Document image retrieval through word shape coding", ***Pattern Analysis and Machine Intelligence, IEEE Transactions on***, Vol. 30, No. 11, (2008), 1913-1918.
10. Zagoris, K., Papamarkos, N. and Chamzas, C., "Web document image retrieval system based on word spotting", in Image Processing, 2006 IEEE International Conference on, IEEE. (2006), 477-480.
11. Keyvanpour, M. and Tavoli, R., "Feature weighting for improving document image retrieval system performance", ***IJCSI International Journal of Computer Science Issues***, Vol. 9, No. 3, (2012), 125-130..
12. MacArthur, S. D., Brodley, C. E., Kak, A. C. and Broderick, L. S., "Interactive content-based image retrieval using relevance feedback", ***Computer Vision and Image Understanding***, Vol. 88, No. 2, (2002), 55-75.
13. Zhou, X. S. and Huang, T. S., "Relevance feedback in image retrieval: A comprehensive review", ***Multimedia Systems***, Vol. 8, No. 6, (2003), 536-544.
14. Rota Bulò, S., Rabbi, M. and Pelillo, M., "Content-based image retrieval with relevance feedback using random walks", ***Pattern Recognition***, Vol. 44, No. 9, (2011), 2109-2122.
15. Su, Z., Zhang, H., Li, S. and Ma, S., "Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning", ***Image Processing, IEEE Transactions on***, Vol. 12, No. 8, (2003), 924-937.
16. Manning, C. D., Raghavan, P. and Schütze, H., "Introduction to information retrieval", Cambridge University Press Cambridge, Vol. 1, (2008).
17. Tan, C. L., Huang, W., Yu, Z. and Xu, Y., "Imaged document text retrieval without OCR", ***Pattern Analysis and Machine Intelligence, IEEE Transactions on***, Vol. 24, No. 6, (2002), 838-844.
18. Keyvanpour, M. and Moghadam Charkari, N., "Interactive retrieval of natural images using multiple instance learning", ***Journal of Iranian Association of Electrical and Electronics Engineers***, Vol. 6, No. 1, (2009) 19-35. .

Document Image Retrieval Based on Keyword Spotting Using Relevance Feedback

M. Keyvanpour [a], R. Tavoli[b], S. Mozaffari[c]

a Department of Computer Engineering Alzahra University, Tehran, Iran
b Department of Electrical and Computer, Islamic Azad University, Qazvin Branch, Qazvin, Iran
c Electrical and Computer Engineering Departmet, Semnan University, Semnan, Iran

P A P E R I N F O

Paper history:
Received 15 March 2013
Received in revised form 11 June 2013
Accepted 20 June 2013

Keywords:
Relevance Feedback
Document Image
Information Retrieval
Keyword Spotting

چکیده

کشف کلمه کلیدی یکی از روش‌های معروف در بازیابی تصاویر اسناد است. در این روش، جستجو در مجموعه‌ای از تصاویر اسناد بر اساس پرس و جوی تصویر کلمه صورت می‌گیرد. در این مقاله یک روش برای بازیابی تصاویر اسناد مبتنی بر کلمه‌ی کلیدی پیشنهاد شده است. در روش پیشنهادی، یک معماری با استفاده از بازخورد مرتبط ارائه می‌شود. بازخورد مرتبط, یک روش تعاملی و موثر است که کارایی سیستم بازیابی تصاویر اسناد را افزایش می‌دهد. در روش پیشنهادی ما چندین استراتژی از بازخوردهای مثبت و منفی شامل "تنها بازخورد مثبت", "تنها بازخورد منفی" و "بازخورد مثبت ومنفی" را با هم مقایسه می‌نماییم. نتایج نشان می‌دهد که با استفاده از بازخورد مرتبط در بازیابی تصاویر اسناد کارایی بهتری نسبت به بازیابی تصاویر اسناد معمولی بدست می‌آید.

doi: *10.5829/idosi.ije.2014.27.01a.02*