2012 10th IAPR International Workshop on Document Analysis Systems

Efficient word retrieval using a multiple ranking combination scheme

G. Louloudis[1], A. L. Kesidis[2,1] and B. Gatos[1]

[1]Computational Intelligence Laboratory
Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos"
GR-15310 Athens, Greece
{louloud, akesidis, bgat }@iit.demokritos.gr

[2]Department of Surveying Engineering
Technological Educational Institution of Athens
GR-12210 Athens, Greece
akesidis@teiath.gr

***Abstract*—Word retrieval is an important task in the area of document analysis and recognition. The selection of appropriate features is a crucial step in the word matching and retrieval process. Several efficient techniques have been proposed which use a wide range of features. This paper proposes a methodology for the efficient fusion of multiple ranking results produced by different word matching techniques. Specifically, a Minimum Ranking method is proposed for the combination of two or more ranking results. The method is compared with two state-of-the-art ranking fusion methods. The experimental results show that the fusion of the ranked results outperforms the ranking efficiency of the individual systems. Moreover, the proposed Minimum Ranking method outperforms the other two state-of-the-art fusion methods.**

Keywords - word retrieval; multiple rankings; ranking combination; data fusion; word matching

I. INTRODUCTION

Document analysis and recognition is the scientific area which aims to the extraction of information from document images. Nowadays, due to the mass digitization of the historical content of libraries around the world, there is a growing need for systems and tools that will automatically provide ways of making the content which is hidden in the document images publicly available. One of the issues concerning the field of document analysis and recognition is the recognition of text in historical documents. Optical character recognition (OCR) systems which are responsible for producing the correct transcription contained on historical document images often produce erroneous results due to document degradations as well as imperfect character segmentation.

Word spotting is a content-based retrieval procedure that spots words directly on document images with the help of efficient word matching while avoiding conventional OCR procedures [1], [2]. We will refer to word spotting as word retrieval since the spotting of words on document collections is a retrieval procedure. The input to a word retrieval system is a word query. The word query comprises either an actual example from the collection of interest or it is artificially generated from an ASCII keyword. The output of a word retrieval system is a list of word images which are ranked according to the degree of similarity compared to the word query. Ideally, all the instances of the word query which appear in the collection will be placed on the top ranking positions.

In this paper, we propose a methodology for the efficient fusion of multiple ranking results produced by different word matching techniques. Data fusion is generally defined as the use of techniques that combine data from multiple sources or systems. In the information retrieval area, fusion corresponds to the merging of retrieval results of multiple systems. The input of a data fusion algorithm is a set of ranked lists and the result is a single ranked list aiming to provide improved retrieval efficiency. In the proposed work, it is shown that fusion of ranked lists from two word retrieval systems leads to improved accuracy. The remainder of this paper is organized as follows: Section II describes related work on word spotting and data fusion. In Section III the proposed methodology is detailed. Conclusions and future work plans are provided in section IV.

II. RELATED WORK

In the literature, word spotting appears under two distinct trends: the segmentation-based approach and the segmentation-free approach. In the segmentation-based approach, the word segmentation stage is mandatory in order to produce word candidates that will be matched with the word query [2], [3], [4], [5]. Concerning the segmentation-free approach, the query word image is fitted to the corresponding word images in the document without any segmentation involved [6], [7], [8]. Detailed surveys on document indexing and retrieval can be found in [1], [9].

Data fusion is mainly used in the area of information retrieval. In general, a data fusion algorithm accepts two or more ranked lists and merges them into a single ranked list with the aim of improved retrieval efficiency [10]. There are mainly two categories of data fusion techniques: (i) methodologies that use the similarity values from each ranked list in order to produce the final ranking list and (ii) methodologies which use the ranking information from each list in order to create the final ranking. Known methods in the first category can be found in [11]. Concerning the second category of methodologies these include the Rank Position, the Borda Count and the Condorcet methods [10].

To our knowledge, no previous use of data fusion methodologies was used in the area of word spotting. In this paper, we apply several data fusion techniques on the results of two word spotting systems. The experimental results show

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.34

that data fusion on the results of individual word spotting techniques improves the performance compared to the individual systems.

III. PROPOSED METHODOLOGY

In a typical segmentation based word spotting system, the document image is first segmented into words and then the following main processes are applied: a) word preprocessing, b) feature extraction and c) word matching. The proposed methodology combines two word spotting systems which share a common word preprocessing and word matching framework, while differ on the feature extraction stage. At the end, a data fusion stage is introduced in order to produce the final ranking list. Figure 5 depicts the system architecture.

A. Word Preprocessing

The word preprocessing consists of three distinct steps: i) noise removal and image enhancement, ii) slant correction and iii) word normalization. Noise removal and image enhancement is accomplished using a 3x3 median filter. The slant correction procedure is described in [12]. Finally, the word normalization is based on placing the baselines (upper and lower) of the words on the center of the matrix [13], [14]. The final word image dimensions are 300x90. Figure 1 illustrates a word image example after applying each of the preprocessing steps.

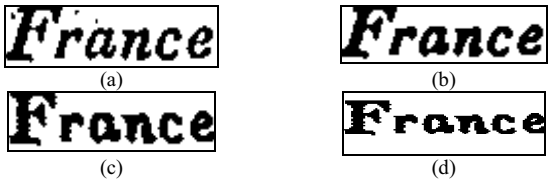


Figure 1. Preprocessing steps on a word image example: (a) initial image, (b) image after noise removal and enhancement, (c) image after slant correction and (d) image after size normalization.

B. Feature Extraction

Features based on zones are of the most popular and efficient statistical features and provide high computational speed and low complexity for character and word recognition. They are calculated by the density of pixels or pattern characteristics in several zones in which we divide the pattern frame. In particular, standard zoning methods are defined according to a $N \times M$ regular grid superimposed on the image body [15]. Zoning features are calculated directly on a size normalized image I as follows:

$$denF_{nm} = \frac{1}{K * \Lambda} \sum_{x=(n-1)K}^{nK-1} \sum_{y=(m-1)\Lambda}^{m\Lambda-1} I(x,y) \quad (1)$$

where K, Λ correspond to the width and height of the window and n=1..N and m=1..M.

We chose two variations of zoning features in order to produce two individual word spotting systems. Concerning the first variation of zoning features, we conducted an experiment of several window sizes with respect to the width and height of the window. The dataset of the experiment is detailed in Section IV. The final values of K and Λ correspond to the window's width and height which produce the best word retrieval performance. Table I contains the retrieval performance with respect to different width and height of the window. As is it observed the width and height which produced the best result are 25 and 6, respectively.

TABLE I. WORD RETRIEVAL RESULTS FOR SEVERAL WINDOW SIZES.

		Height							
		3	***5***	***6***	***9***	***10***	***15***	***18***	***30***
Width	***3***	85,0	85,1	84,9	84,0	83,6	82,5	81,1	79,6
	4	85,7	85,3	85,0	84,3	84,0	82,7	81,4	80,1
	5	86,0	85,7	85,5	84,9	84,7	83,1	81,9	80,5
	6	86,3	86,1	85,9	85,0	84,8	83,6	82,3	81,1
	10	87,1	87,0	86,8	85,9	85,7	84,7	83,2	82,4
	12	87,5	87,7	87,7	86,6	86,1	85,7	83,3	81,9
	15	88,7	88,4	88,5	87,0	86,3	84,9	82,6	79,2
	20	86,7	86,6	87,1	85,4	83,2	81,7	77,9	72,1
	25	88,6	88,9	90,1	89,1	86,2	86,5	80,3	76,1
	30	83,9	84,7	85,8	84,3	81,0	83,7	72,6	70,4
	50	70,2	72,3	75,7	73,0	70,0	76,4	61,2	59,3
	60	61,3	65,6	68,2	66,0	59,5	64,7	50,4	47,4

The second variation of zoning features is described in [13]. These zones are adaptive in the sense that the position of each zone is adjusted based on local pattern information. Specifically, this adjustment is performed by moving every zone towards the pattern body. The horizontal and vertical range for adjusting the position of the zones is defined by parameters λ_x and λ_y. The offset that is used for adjusting the zone position is calculated by maximizing the local pixel density around the zone. For our methodology we used λ_x= λ_y=4 whereas the window size is 10x10. These parameters produced the best results in [13]. Figure 2 presents the preprocessed word image of Figure 1 superimposed with the fixed position windows. The positions of the corresponding adaptive windows are presented in Figure 3.

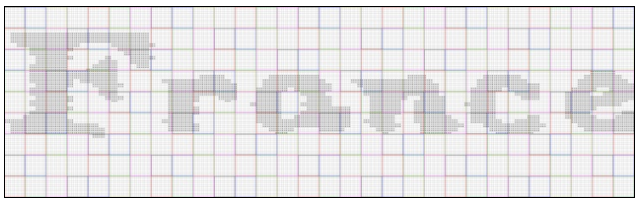


Figure 2. Zoning procedure example.

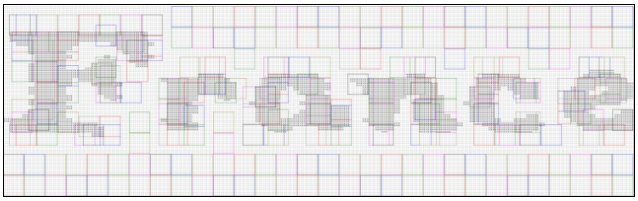


Figure 3. Posistion adjustment for all zones in the image of Figure 2.

C. Word Matching

The retrieval result of each word spotting system is based on the calculation of the Euclidean distance between the features of the word query and each word image of the dataset. If $p=(p_1,p_2,...,p_n)$ and $q=(q_1,q_2,...,q_n)$ are two feature vectors, the Euclidean distance $d(p,q)$ is defined as:

$$d(p,q)=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2} \tag{2}$$

The ranked list is produced by sorting the results of the word matching from the most similar to the less similar.

D. Data Fusion

This is the final stage of the proposed word retrieval system that uses the ranked lists produced by the two individual systems and produces the final ranked result list. Three different methods are examined all of which are based on the combination of the ranking lists. These include: i) the Rank Position (reciprocal rank) method [10], ii) the Borda Count method [10] and iii) the proposed Minimum Ranking method. More detailed description of these methods is given in the next paragraphs.

Rank position method. In order to merge the results of the word retrieval systems into a final list only the rank positions of the corresponding words are used. The following equation shows the computation of the final ranking score of word *i*, *i*=1..*N* using the position information of this word in all systems *j*.

$$r(d_i)=\frac{1}{\sum_j 1/\text{position}(d_{ij})} \tag{3}$$

Borda Count method. This method is based on democratic election strategies. According to this method, the highest ranked word in a system gets *N* Borda points and each subsequent gets one point less where *N* is the size of the ranked list.

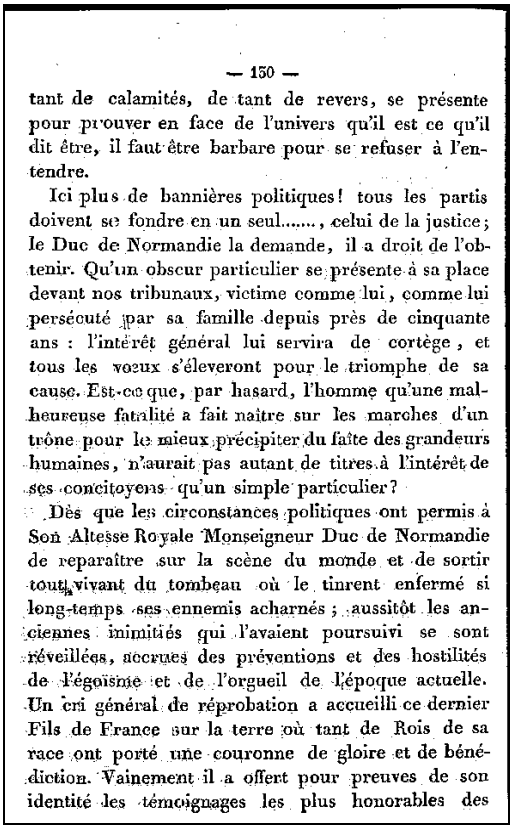
$$r(d_i)=\sum_j(N-\text{position}(d_{ij})) \tag{4}$$

Minimum Ranking method. For each retrieved word we find the minimum rank position on all ranked lists and this value is considered as the distance measure.

$$r(d_i)=\min_j(\text{position}(d_{ij})) \tag{5}$$

IV. Evaluation and Experimental Results

The proposed methodology was tested on a historical French book [16] which contains 153 pages and 46197 words. A sample page of the French book is shown in Figure 4. As it was described in the previous section no word segmentation was applied since the word segmentation as well as the ASCII ground truth were manually created. We randomly selected five instances of the words 'France', 'Louis', 'famille', 'mort' and 'justice', thus yielding 25 queries in total. The total number of instances of those words in the document corpus is 44, 156, 47, 51 and 44, respectively.



— 150 —

tant de calamités, de tant de revers, se présente pour prouver en face de l'univers qu'il est ce qu'il dit être, il faut être barbare pour se refuser à l'entendre.

Ici plus de bannières politiques! tous les partis doivent se fondre en un seul....., celui de la justice; le Duc de Normandie la demande, il a droit de l'obtenir. Qu'un obscur particulier se présente à sa place devant nos tribunaux, victime comme lui, comme lui persécuté par sa famille depuis près de cinquante ans : l'intérêt général lui servira de cortège, et tous les voeux s'éleveront pour le triomphe de sa cause. Est-ce que, par hasard, l'homme qu'une malheureuse fatalité a fait naître sur les marches d'un trône pour le mieux précipiter du faîte des grandeurs humaines, n'aurait pas autant de titres à l'intérêt de ses concitoyens qu'un simple particulier?

Dès que les circonstances politiques ont permis à Son Altesse Royale Monseigneur Duc de Normandie de reparaître sur la scène du monde et de sortir tout vivant du tombeau où le tinrent enfermé si long-temps ses ennemis acharnés ; aussitôt les anciennes inimitiés qui l'avaient poursuivi se sont réveillées, accrues des préventions et des hostilités de l'égoïsme et de l'orgueil de l'époque actuelle. Un cri général de réprobation a accueilli ce dernier Fils de France sur la terre où tant de Rois de sa race ont porté une couronne de gloire et de bénédiction. Vainement il a offert pour preuves de son identité les témoignages les plus honorables des

Figure 4. A document image sample.

We conducted the same experiment with [13], [14] in order to be directly comparable. Let *n_inst* be the total number of instances of a word in the ground truth and *n_corr* the number of correct instances of the word in the first *n_inst* retrieved instances. The word retrieval performance (*WRP*) can be calculated as follows:

$$WRP=\frac{n\_corr}{n\_inst} \tag{6}$$

In order to test the proposed methodology, we conducted three experiments which differ on the steps that were applied in the preprocessing stage. The only preprocessing step that was used in the first experiment was word image normalization. Same experimental conditions were also

applied in [13], [14]. Table II contains the word retrieval performance of the two individual systems (BS1 and BS2) as well as the performance obtained after applying the data fusion methods which are described in the previous section. It is clear that all data fusion methodologies outperform the individual systems. Columns *N, M* are defined as follows:

Let N_i be the total number of instances of query word *i* in the ground truth and M_i the number of correct instances of this word in the first N_i retrieved instances, where i=1…25. Then, columns *N* and *M* are given by the following equations:

$$N = \sum_{i=1}^{25} N_i \tag{7}$$

$$M = \sum_{i=1}^{25} M_i \tag{8}$$

TABLE II. WORD RETRIEVAL PERFORMANCE USING SIZE NORMALIZATION

Method	**N**	**M**	***WRP* (%)**
Adaptive Zoning BS1 [13]	1710	1539	90,0%
Fixed Zoning BS2	1710	1541	90,1%
Data Fusion using Rank Position	1710	1574	92,0%
Data Fusion using Borda Count	1710	1565	91,5%
Data Fusion using Min Ranking	1710	1573	92,0%

In the second experiment we applied two preprocessing steps: noise removal and image enhancement as well as word normalization. It is evident that the performance of both systems BS1 and BS2 was improved compared to the previous experiment where we did not use the noise removal and image enhancement step. We should also note that all data fusion methodologies also improved and still outperformed the individual systems. Table III presents the performance of all systems after conducting the second experiment.

TABLE III. WORD RETRIEVAL PERFORMANCE USING SIZE NORMALIZATION AND IMAGE ENHANCEMENT

Method	**N**	**M**	***WRP* (%)**
Adaptive Zoning BS1 [13]	1710	1561	91,3%
Fixed Zoning BS2	1710	1553	90,8%
Data Fusion using Rank Position	1710	1585	92,7%
Data Fusion using Borda Count	1710	1574	92,0%
Data Fusion using Min Ranking	1710	1586	92,8%

In the third experiment we used all preprocessing steps (i.e. noise removal and image enhancement, slant correction and word normalization). The performance of all systems is presented in Table IV. It is depicted that there was still an improvement in all systems. A closer look at the performance of the data fusion methods shows that in almost all experiments the Minimum Ranking method outperforms the other two methods. Also, it can be noticed that the Borda Count method shows a small drop of performance compared with the BS1 method.

TABLE IV. WORD RETRIEVAL PERFORMANCE USING SIZE NORMALIZATION, IMAGE ENHANCEMENT AND SLANT CORRECTION

Method	**N**	**M**	***WRP* (%)**
Adaptive Zoning BS1 [13]	1710	1583	92,6%
Fixed Zoning BS2	1710	1561	91,3%
Data Fusion using Rank Position	1710	1591	93,0%
Data Fusion using Borda Count	1710	1581	92,5%
Data Fusion using Min Ranking	1710	1593	93,2%

V. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a methodology for the efficient fusion of multiple results produced by different word matching techniques. Three different methods are examined all of which are based on the combination of the ranking lists. These include the Rank Position method, the Borda Count method and the proposed Minimum Ranking method. The experimental results show that in almost all cases the fusion of the ranked results outperforms the ranking efficiency of the individual systems. Moreover, in all experiments the proposed Minimum Ranking method outperforms the other two fusion methods.

Future work includes the investigation of the data fusion process when more than two word spotting systems are used. Furthermore, the fusion applicability should be investigated in case where word spotting systems using different types of features are combined.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement n° 215064 (project IMPACT).

REFERENCES

[1] A. Murugappan, B. Ramachandran, P. Dhavachelvan, “A survey of keyword spotting techniques for printed document images”, Artificial Intelligence Review, Vol. 35, No 2, pp. 119-136, 2011.

[2] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis and S. J. Perantonis, "Keyword-guided word spotting in historical printed documents using synthetic data and user feedback", International Journal on Document Analysis and Recognition (IJDAR), special issue on historical documents, Vol. 9, No. 2-4, pp. 167-177, 2007.

[3] T. M. Rath, and R. Manmatha, “Features for word spotting in historical documents” in Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03), pp 218-222, Edinburg, Scotland, August 2003.

[4] A. Balasubramanian, M. Meshesha, and C. V. Jawahar “Retrieval form document image collections”, in Proceedings of the 7th International Workshop on Document Analysis Systems (DAS’06), pp 1-12, Nelson, New Zealand, February 2006.

[5] A. Bhardwaj, D. Jose and V. Govindaraju, “Script independent word spotting in multilingual documents”, In Proceedings of the 2nd International Workshop on Cross Lingual Information Access (CLIA’08), pp. 48-54, Hyderabad, India, January 2008.

[6] B. Gatos and I. Pratikakis, “Segmentation-free word spotting in historical printed documents,” in Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR’09), pp. 271–275, Barcelona, Spain, July 2009.

[7] Y. Leydier, A. Ouji, F. LeBourgeois, and H. Emptoz, “Towards an omnilingual word retrieval system for ancient manuscripts,” Pattern Recognition, vol. 42, no. 9, pp. 2089– 2105, 2009.

[8] M. Rusiñol, D. Aldavert, R. Toledo and J. Lladós, “Browsing heterogeneous document collections by a segmentation-free word spotting method ,” in Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR’11), pp. 63-67, Beijing, China, September 2011.

[9] D. Doermann, “The indexing and retrieval of document images: A survey,” Computer Vision and Image Understanding, vol. 70, pp. 287–298, 1998.

[10] R. Nuray and F. Can, “Automatic ranking of information retrieval systems using data fusion,” Internation Journal of Information Processing and Management, Vol. 42, Iss. 3, pp.595-614, 2006.

[11] E.A. Fox and J.A. Shaw, “Combination of multiple searches,” in Proc. Of the 2nd Text REtrieval Conference (TREC-2), National Institute of Standards and Technology Spercial Publication 500-215, pp.243-252, 1994.

[12] A. Vinciarelli and J. Luettin, “A new normalization technique for cursive handwritten words,” Pattern Recognition Letters,Vol 22, no. 9, pp 1043-1050, 2001.

[13] B. Gatos, A. L. Kesidis and A. Papandreou, “Adaptive zoning features for character and word recognition,” in Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR’11), pp. 1160-1164, Beijing, China, September 2011.

[14] S. Colutto and B. Gatos, “Efficient word recognition using a pixel-based dissimilarity measure,” in Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR’11), pp. 1110-1114, Beijing, China, September 2011.

[15] M. Bokser , “Omnidocument technologies”, Special Issue on Optical Character Recognition, Proceedings of the IEEE, Vol. 80, pp. 1066-1078, 1992.

[16] Le Dernier fils de France, ou le Duc de Normandie, fils de Louis XVI et de Marie-Antoinette, par A. Bibliothèque nationale de France, 1838.

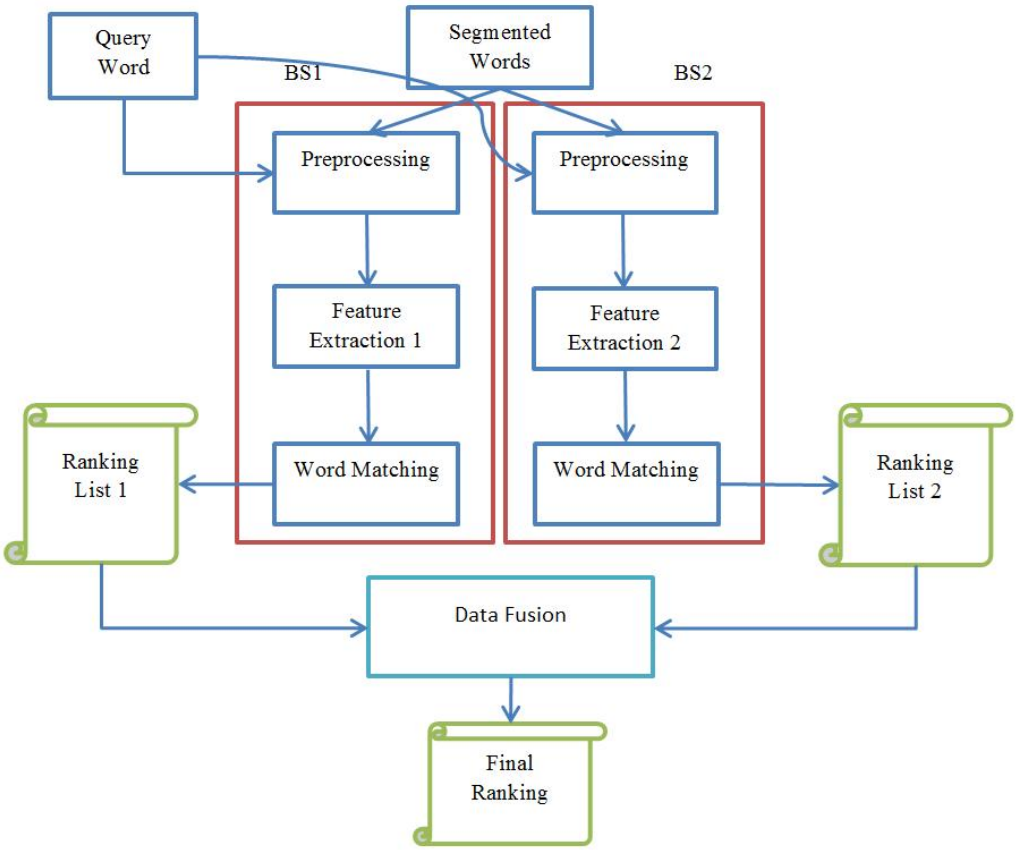


Figure 5. The overall system architecture.