한글 단어의 고속 검색을 위한 두 단계 알고리즘

양진호[0], 오일석
전북대학교 컴퓨터과학과
jhyang@cs.chonbuk.ac.kr, isoh@moak.chonbuk.ac.kr

A Two-Pass Algorithm for Fast Retrieval of Korean Words

Jin-Ho Yang, Il-Seok Oh

Department of Computer Science, Chonbuk National University

요 약

대용량 한글 문서를 대상으로 하는 검색 시스템은 고속의 단어 매칭 알고리즘을 필요로 한다. 이 논문은 두 단계 매칭 접근 방법을 제시하고 그 유용성을 실험을 통해 입증한다. 특징으로는 웨이브렛(wavelet) 계수를 사용하여 첫 단계에서는 적은 수의 특징만을 사용하여, 거친 정합(coarse matching)을 하며 두 번째 단계에서는 미세한 정합(fine matching)을 한다. 거친 정합은 가능성이 없는 단어를 아주 빠른 속도로 걸러내는 역할을 한다. 실제 한글 단어 영상 데이터베이스에 적용한 실험 결과 검색률의 희생 없이 약 7배의 속도 향상을 얻었다.

1. 서 론

컴퓨터의 발전과 인터넷이 보편화되면서 정보 수집과 지식 전달 매체는 급격히 전환되고 있다. 인쇄된 문서나 도서가 아닌 디지털 도서관이나 광 파일 시스템 같은 데이터베이스를 통해 사용자는 정보를 검색하고 수집한다. 이러한 시스템을 구축하기 위한 방법으로 수십만 내지 수백만 건의 문서를 저장하는 작업은 다양하게 이루어지고 있다.

이러한 데이터베이스를 구축하기 위해 전문가의 수작업을 통해서 문서를 입력하는 방법은 많은 노동력과 비용을 필요로 한다[1]. 이런 단점을 줄이기 위해 인쇄된 문서를 OCR(Optical Character Reader)을 이용한 문서 입력 방법을 사용하는 연구[2]도 진행되고 있으나 이 접근 방법은 오인식 문제로 인한 한계를 갖고 있다. 또 다른 방법은 영상-기반 문서 입력 방법이 있다. 이 방법은 스캔한 문서를 영상 자체로써 저장한다[3]. 예를 들어 KISTI(한국과학기술연구정보원)에서는 각종 저널들을 스캔하여 영상 형태로 저장하고 이를 기반으로 원문 검색을 해주고 있다. 이렇게 입력된 방대한 문서들을 검색하기 위해서는 초고속의 매칭 알고리즘이 필수적이다. 영상-기반 문서 검색 방법이 한글 단어 검색에 적용된 사례는 김혜금 논문에서 웨이브렛(wavelet) 특징을 이용한 검색 방법을 찾아볼 수 있다[4].

본 논문은 인쇄된 한글 문서에서 단어 검색을 위해 웨이브렛에 기반한 두 단계 검색 방법을 제시한다. 스캔 받은 문서 영상은 전처리 단계를 거쳐 단어 단위로 분할(segmentation) 되었다는 가정아래 웨이브렛 변환 방법으로 낱자별로 특징을 추출한다. 추출된 특징들은 매칭 알고리즘을 적용하여 매칭 점수를 계산한다. 계산된 값은 검색 조건에 의하여 검색률과 속도를 측정하였다. 이와 같이 각 단어에 적용되는 특징들은 웨이브렛 계수가 많을수록 검색률은 좋아지고 속도는 떨어지는 원리를 알 수 있다. 만약 대용량 문서에 적용하자면 아무리 좋은 검색률 일지라도 초고속으로 처리하지 못하면 문서 검색에 응용하기 힘들 것이다. 따라서 검색률이 떨어지지 않는 조건아래 초고속으로 단어 검색을 위한 방법을 추구한다.

첫 단계에서는 적은 수의 웨이브렛 계수를 사용하여 거친 정합(coarse matching)을 하고 두 번째 단계에서는 미세한 정합(fine matching)을 한다. 거친 정합은 수많은 단어들 중에서 가능성이 없는 단어를 아주 빠른 속도로 제거함으로서 두 번째 단계에 사용할 단어의 수를 줄임을 목적으로 한다. 두 번째 단계의 역할은 검색률을 가장 높일 수 있는 적절한 수의 계수를 선택하여 축소된 단어 데이터베이스에 미세한 정합을 적용한다. 이렇게 두 단계 검색 방법을 통해 좀 더 빠른 실험 결과를 얻을 수 있었다.

2. 특징 추출

하르 웨이브렛(Harr wavelet)은 정직교성(orthonormality)과 빠른 계산 능력 등의 좋은 특성들 때문에 영상 압축 [5]과 영상 검색[6]에 널리 사용되어 왔다. 본 논문에서도 하르 웨이브렛 변환(Haar wavelet transform)을 이용하여 단어에 대한 낱자별 특징들을 추출하였다. 이렇게 단어 영상에서 추출된 특징들은 데이터베이스에 저장되고 사용자가 검색할 질의어도 같은 방식으로 특징을 추출하여 매칭 하는데 사용한다. 예를 들어, 32*32 크기의 단어 영상에 대한 압축 계수열은 총 1024개의 특징 값을 얻는다. 추출된 압축 계수열은 1024개중 선택적으로 매칭 하는데 사용된다.

3. 매칭 알고리즘

매칭 알고리즘은 문자 대 문자로 수행하므로, 검색 대상 단어와 질의 단어로 생각하면 된다. 검색 대상이 되는 단어는 목적단어 T로 표기하고 사용자가 입력한 질의 단어는 Q라고 표기한다. T와 Q는 $T(C^t_1, C^t_2, C^t_3, \cdots, C^t_N)$와 $Q(C^q_1, C^q_2, C^q_3, \cdots, C^q_N)$로 표현할 수 있다. 여기서 C는 단어의 문자이고, N은 단어를 형성하는 문자의 갯수이다. 하르 웨이브렛을 통해 추출된 문자별 특징들은 압축 계수 목록 $L^t_{compressed}$와 $L^q_{ordered}$로 표기하며 다음과 같이 쓸 수 있다.

$L^t_{compressed} = (\alpha^t, \omega^t_1, \omega^t_2, \ldots\ldots, \omega^t_K)$ 와

$L^q_{ordered} = (\alpha^q, \omega^q_{\pi(1)}, \omega^q_{\pi(2)}, \ldots\ldots, \omega^q_{\pi(K)})$, $\pi(i)$는 내림차순으로 정렬된 색인들을 가리킴.

매칭 대상이 되는 문자 $L^t_{compressed}$는 추출된 상태로 저장하고 질의 문자는 $L^q_{ordered}$는 내림차순으로 정렬하여 저장한다. 저장된 압축 계수들은 큰 값일수록 원본 문자 영상을 잘 표현하므

로 작은 값들을 제거해도 정보의 손실은 최소로 줄이면서 매칭에 사용할 수 있다. 단어의 특징에 따라 계수 K개의 선택은 매칭 알고리즘의 성능과 속도에 영향을 주므로 질의 문자에서 내림차순으로 정렬된 계수들중 상위 K개만 선택하여 매칭 공식에 사용한다. 즉, C^t_i에서 $L^t_{compressed}$, C^q_i에서 $L^q_{ordered}$를 얻은 후 두 문자 사이의 매칭 점수를 계산하기 위해 아래 공식을 사용한다.

$$MS(L^t, L^q) = \Sigma_{i=1,K} |\omega^t_{\pi(i)} - \omega^q_{\pi(i)}| + h*|\alpha^t - \alpha^q|$$

위에서 제시한 문자 대 문자 매칭 공식을 이용하여 매칭 점수에 임계값을 적용해서 판별 여부를 결정한다. 단어 매칭 조건은 $MS(L^t, L^q)$가 임계값 ε_1보다 작고 동시에 문자 N개의 매칭 점수들의 평균이 임계값 ε_2보다 작을 때 단어는 성공적으로 매칭 되었다고 한다.

4. 두 단계 검색 방법

두 단계 검색 방법은 사용자가 질의 단어를 입력하면 단어 영상 데이터베이스에서 거친 정합과 미세한 정합을 하여 검색 결과를 보여준다. 그림 1은 두 단계 검색 방법을 보여 주며, 거친 정합을 하는 부분은 1-단계, 미세한 정합을 2-단계로 구분한다.

1-단계는 초고속 단어 매칭을 목적으로 한다. 적은 계수(K)를 선택하여 속도를 높이고, 재현율을 100%에 가깝게 임계값을 적용하여 찾을 단어의 손실 없이 후보 단어들을 구성한다.

2-단계는 정확율을 목적으로 단어 매칭을 한다. 1-단계를 통과한 후보 단어들을 대상으로 적당한 계수 선택과 임계값을 적용하여 매칭 결과를 보여준다. 2-단계의 계수 선택과 임계값 적용은 정확율을 높이는 결정적인 역할을 한다.

이렇게 두 단계 검색 방법을 사용하므로 전체적인 검색률은 유지하며 속도는 높일 수 있다. 예를 들어 100만 단어 중 찾아야 할 단어가 100단어가 있다고 가정하자. 1-단계에서 빠른 검색으로 1만 단어정도 간추려서 후보 단어를 구성하면 2-단계에서는 후보 단어 1만개에서 100단어를 검색하는 방법이다. 그림 1에서 W는 전체 단어 영상이고 X는 후보 단어 영상을 말하며 Y는 검색된 최종 단어 영상으로 표기하였다.

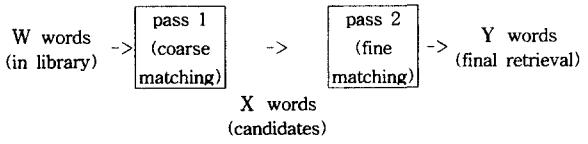


그림 1. 두 단계 검색

5. 실험 결과

5.1 실험 환경

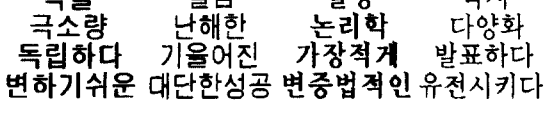
제안한 두 단계 검색 방법을 테스트하기 위해 전남대에서 구축한 단어를 목적단어(Target word)로 사용하였고, 질의 단어(Query word)는 한글 워드 프로세서를 사용하여 2,350개의 한글 문자 영상을 화면상에 생성하고, 영상 편집 소프트웨어를 사용하여 문자 영상을 편집하였다.

실험에 사용한 단어의 종류는 고딕체와 명조체 각각에 대해 bold와 regular 속성을 가지고 있고 폰트 크기는 10-14 points 고 2-5글자로 다양하다. 표 1과 그림 2는 데이터베이스의 단어 영상 예제를 보여준다.

제안한 두 단계 검색의 성능과 속도 평가를 위해 1-단계만 적용한 결과를 비교 대상으로 삼았고, 펜티엄III 800Mhz, 128MB 메인 메모리의 PC에서 테스트하였다. 검색 속도의 차이는 PC의 사양에 따라 다를 수 있다.

표 1. 목적 단어로 사용한 단어 영상 데이터 베이스

폰트 종류	속성	단어 영상 갯수
고딕체	bold	1,175
	regular	1,115
명조체	bold	1,183
	regular	1,098



고딕bold 고딕regular 명조bold 명조regular

그림 2. 실험에 사용한 한글 단어 영상 예

5.2 성능 평가

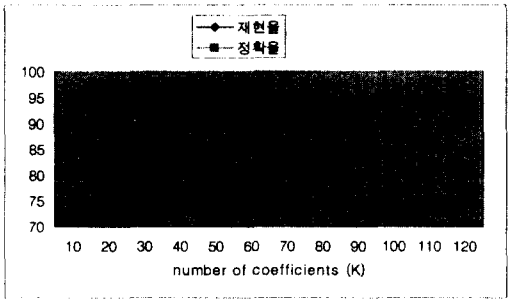
표 2와 그림 3은 1-단계만 적용한 결과를 보여준다. 매칭 알고리즘을 1-단계만 적용하여 특징 계수(K)를 10부터 120까지 측정하였다. 30개 계수를 선택했을 경우 가장 높은 검색률을 보여 주지만, 적은 계수를 선택했을 때 높은 검색 속도를 얻을 수 있었다.

표 3과 그림 4는 본 논문에서 제안한 2-단계 검색 방법을 사용하여 실험한 결과를 보여준다. 1-단계만 적용한 결과를 바탕으로 계수를 선택하였고 초고속 검색을 목적으로 실험한 결과이다.

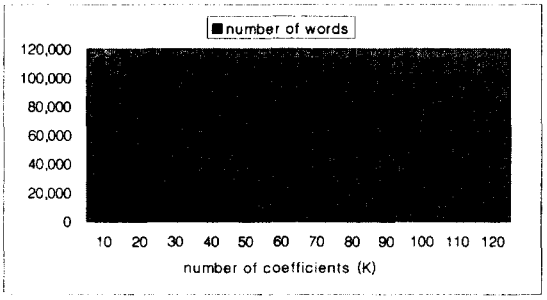
두 단계 검색 방법에서 1-단계에 적용된 계수는 10부터 하나씩 줄여가며 적용하였고, 2-단계에 적용된 계수는 30개로 고정시켜 실험하였다. 30개의 계수를 선택한 이유는 1-단계만 적용했을때 정확율이 가장 높았기 때문이다. 두 실험 결과를 비교해 보면 비슷한 정확율 약 96%를 보이는 단어 검색 속도를 보면 33,063.49 대 236,225.32의 속도대비를 알 수 있다. 즉, 비슷한 검색률에서는 최고 약 7배까지 빠른 검색 속도를 보이고 있다.

표 2. 1-단계만 적용한 검색 성능 표

K	Recall ratio(%)	Precision ratio(%)	Speed (words/second)
10	95.21	92.38	96,559.40
20	96.47	95.62	48,934.93
30	96.99	96.74	33,063.49
40	96.89	96.71	25,682.16
50	96.40	95.99	20,019.01
60	95.69	94.88	17,560.56
70	94.50	93.69	14,808.29
80	95.18	93.90	12,708.94
90	95.03	93.70	11,493.38
100	95.04	94.02	10,394.26
110	95.61	94.31	9,465.29
120	95.45	94.47	9,015.65



(a) 1-단계 적용한 검색 성능 그래프

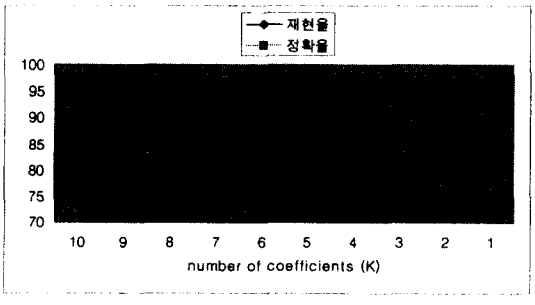


(b) 초당 검색 속도 그래프

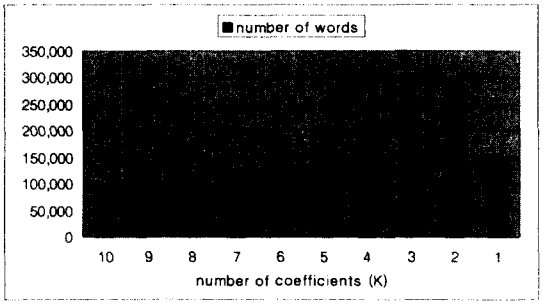
그림 3. 1-단계만 적용한 검색 실험 결과

표 3. 2-단계 적용한 검색 성능 표

K (1-단계)	K (2-단계)	Recall ratio(%)	Precision ratio(%)	Speed (words/second)
10	30	97.18	96.73	85,610.28
9	30	97.20	96.74	96,497.63
8	30	97.18	96.74	106,213.42
7	30	97.20	96.77	112,703.88
6	30	97.26	96.77	130,553.27
5	30	97.30	96.76	156,891.80
4	30	97.20	96.77	199,918.37
3	30	96.87	96.40	236,225.32
2	30	95.36	94.98	286,513.68
1	30	87.48	87.25	147,213.75



(a) 2-단계 검색 성능 그래프



(b) 초당 검색 속도 그래프

그림 4. 2-단계 검색 실험 결과

실험 결과 압축 계수의 선택에 따라 검색률과 속도를 향상시킬 수 있으며, 1-단계에서 계수를 적게 선택할 수록 정확율은 유지하면서 빠른 단어 검색 속도를 보여주고 있다. 그렇지만 최소의 계수 선택은 정확율이 급격히 떨어지는 결과를 얻을 수 있다.

웨이브렛으로 압축된 단어 영상을 한번에 검색하는 것보다 두 번에 걸쳐 검색하는 것이 좀더 빠른 검색 속도를 보였고 전체 검색 시스템에서 검색 성능을 높이는 최적의 계수 선택과 임계값의 적용이 가장 중요하게 작용함을 알 수 있다.

6. 결론

본 논문에서는 웨이브렛에 기반한 두 단계 한글 단어 검색 방법을 제안하였다. 실험결과 제안한 두 단계 검색 알고리즘은 빠른 속도, 저장공간의 효율성, 신뢰할 수 있는 검색 성능을 제공해 주기 때문에 디지털 도서관이나 광 파일 시스템에서 유용하게 사용할 수 있으리라 생각한다. 또한 제안한 방법은 정확률과 재현율 간의 tradeoff와 검색 성능과 계산 시간간의 tradeoff 조절이 가능하다는 장점을 갖고 있다.

참고 문헌

[1] 김태수, 유양근, 정준민, 최석두, 디지털 도서관, 사이텍 미디어, 2000.

[2] 안재철, 오일석, "문자 인식을 이용한 한글 문서 검색," 한국정보과학회 춘계 학술발표논문집, 제28권, 제1호, pp.544-546, 2001.

[3] D. Doermann, The retrieval of document images: a brief survey, *Proceedings of ICDAR97,* Ulm, pp. 945-949, 1997.

[4] 김혜금, 양진호, 이진선, 오일석, "웨이브렛을 이용한 영상 기반 인쇄 한글 단어 검색," 정보 과학회 논문지, Vol. 28, No. 2, pp.91-103, 2001.

[5] R.A. DeVore, B. Jawerth, and B.J. Lucier, Image compression through wavelet transform coding, *IEEE Trans. on Information Theory,* Vol.38, No.2, pp.719-746, March 1992.

[6] C.E. Jacobs, A. Finkelstein, and D.H. Salesin, Fast multiresolution image querying, *Proceedings of SIGGRAPH95,* pp.277-286, 1995.