한글 초성을 이용한 원문보호 탐색기법

김성환*, 박선영*, 조환규*
*부산대학교 컴퓨터공학과
e-mail:sunghwan@pusan.ac.kr

A Text-Secure Searching Method Using the First Phonemes in Korean Documents

Sung-Hwan Kim*, Sun-Young Park*, Hwan-Gue Cho*
*Dept of Computer Science and Engineering, Pusan National University

요 약

인터넷의 발달과 활발한 지적 활동으로 인하여 정보 교류의 속도와 양이 급격하게 증가하였고 이에 따라 다량의 유사한 정보들 가운데 사용자가 원하는 정보를 탐색하기 위한 다양한 기법들이 연구되어 왔다. 하지만 이러한 탐색 기법을 적용하기에 앞서 탐색 대상이 되는 문서의 원문을 확보하여 데이터베이스를 구축하고 또한 사용자의 질의문을 관리하는 과정에 있어서 사회·법률적인 문제가 방해요소로 작용함에 따라 원문과 질의문의 노출을 최소화하면서도 기존의 정보 탐색 기법들을 적용할 수 있는 공학적 해결 방법이 필요하게 되었다. 본 논문에서는 탐색 대상인 한글 문서의 각 문자를 구성하는 초성만을 추출하여 나열한 문서를 정의하고 이 초성 문서가 원문과 질의문의 노출을 방지하는 동시에 문장 단위 이상의 검색에 있어서 기존의 탐색 방법들이 초성 문서상에서 효과적으로 적용될 수 있음을 실험적으로 입증하였다.

1. 서론

인터넷의 발달과 지적 활동의 증가는 정보의 양산을 불러왔고 이제는 더 이상 정보의 바다라는 용어도 새롭지 않은 사회가 되었다. 이에 따라 다량의 유사 정보들 가운데 사용자의 요구에 적합한 정보를 탐색하기 위하여 색인, 역파일, 확률모델 등을 이용한 다양한 기법들이 연구되어 왔다[1].

기존의 탐색 방법론은 정보 검색 서비스를 제공하기 위하여 먼저 대중에게 공개된 정보를 수집하여 데이터베이스를 구축하는 작업을 필요로 한다. 그러나 타인의 정보를 수집하고 제공하는 과정에서 저작권을 비롯한 많은 법률 분쟁이 발생하고 있다[2]. 특히 전문 기술이나 사적인 정보의 경우에는 저작자가 해당 원문이 불특정다수에게 대가없이 공개되는 것을 원하지 않는 경우가 많기 때문에 각각의 정보들에 대하여 일일이 저작권을 위임받는 것은 매우 어려운 일이다.

또한 고전적인 탐색 기법들은 사용자의 질의문이 그대로 노출되는 것을 피할 수 없기 때문에 사용자가 질의한 내용이 유출되는 경우 사생활 침해나 아이디어 도용과 같은 사회적 문제를 발생시킬 수 있다. 더욱이 이 문제는 시스템 외부의 공격자에 의해서만 발생하는 것이 아니라 시스템 내부의 관련자에 의하여 발생할 수도 있는 문제이기 때문에 방화벽과 같은 네트워크와 관련된 보안 기법으로도 해결할 수 없다.

질의문 노출 문제를 해결하기 위하여 검색자의 질의어를 노출시키지 않으면서 검색을 수행하는 방법들[3]과 데이터베이스 자체를 암호화한 상태로 검색을 수행하는 방법들[4]에 대한 활발한 연구가 진행되고 있으나, 복잡한 연산으로 인한 탐색속도 저하를 극복하기 어려우며 암호화를 하는 경우 암호 키에 관한 안전한 관리 방법이 추가적으로 필요하다. 또한 이 방법들이 해결하고자 하는 문제는 탐색 시스템을 위한 데이터베이스 구축 이후의 문제로 탐색 대상 문서의 원본을 수집해야 한다는 근본적인 문제를 해소하는 것은 불가능하다.

본 논문에서는 탐색 대상 문서의 원문과 질의문의 내용을 검색할 수 있으면서도 노출을 방지할 수 있는 한글 문서 변환 방법으로 한글 초성의 활용 방안을 제안하고 이를 실험적으로 입증하고자 한다.

2. 한글 초성 변환

탐색 대상 문서의 원문과 질의문의 노출을 방지하기 위해서는 다대일 함수가 필요하다. 일반적으로 알려진 해시 함수는 그 설계상의 특성 상 원문이 매우 유사하더라도 그 결과 값이 크게 상이해지므로 유사 문서간의 일관성이 보존되지 않는다. 따라서 다대일 대응을 만족하면서도 유사한 문서 간의 일관성이 보존될 수 있는 방법이 필요하다.

한글의 각 글자는 초성, 중성, 그리고 종성의 세 부분

으로 구성되어 있는데 그 중 초성의 정보량이 가장 높다 [5]. n개의 한글 글자로 이루어진 문서 $D=<x_1,\cdots,x_n>$에 대하여 이 문서를 구성하는 각각의 한글 글자 x_i는 초성 f_i, 중성 m_i, 종성 l_i로 구성되어 있으므로 $x_i=<f_i,m_i,l_i>$와 같이 표현할 수 있다. 한글 글자 집합 X에서 한글 초성 집합 F로 가는 초성 추출 함수 $cho:X\rightarrow F$가 있을 때 한글 초성 변환 $Skin(D)$는 식 (1)과 같이 정의된다.

$$Skin(D)=<cho(x_1),\cdots,cho(x_n)> \quad (1)$$

만약 초성 추출 함수 $cho(x_i)$가 x_i의 초성 f_i만으로 사상된다면 출현 빈도가 낮은 쌍자음 ㄲ,ㄸ,ㅃ,ㅆ,ㅉ 등이 출현하는 경우 원문을 추측할 수 있는 가능성이 높아지게 된다. 따라서 쌍자음의 경우 단자음으로 치환하도록 초성 추출 함수 $cho(x_i)$를 식 (2)와 같이 정의함으로써 쌍자음의 희소 빈도를 이용한 원문 추측을 어렵게 하였다.

$$cho(x_i)=\begin{cases} f_i & \text{if } f_i\text{가 단자음} \\ f_i\text{의 단자음} & \text{if } f_i\text{가 쌍자음} \end{cases} \quad (2)$$

예를 들어 문서 “까치 설날은 어저께고요, 우리 설날은 오늘이래요!”에 대한 변환은 표 1과 같이 나타낼 수 있다.

<표 1> 한글 초성 변환의 예

D	까치설날은어저께고요우리설날은오늘이래요
$Skin(D)$	ㄱㅊㅅㄴㅇㅇㅈㄱㄱㅇㅇㄹㅅㄴㅇㅇㄴㅇㄹㅇ

3. 원문 및 질의문 보호성에 관한 실험

한글 문서로부터 초성만을 추출하는 변환을 통하여 원문 및 질의문에 대한 보호가 가능할 것인지를 실험적으로 증명하고자 한다. 이를 위하여 한글 명사 사전에 수록된 64,261개 명사들의 초성 분포를 구해보았다. 표 2는 한 글자 명사의 초성 분포이며, 표 3은 두 글자 명사의 초성 분포를 나타낸다.

이러한 방식으로 명사 사전에서 각 글자별 초성의 분포를 구한 다음 길이 k의 초성 문자열이 명사의 조합으로 이루어져 있다는 사실이 주어진 경우 복원 가능한 모든 경우의 수를 구해보았다. 예를 들어 길이가 3인 한글 초성 문자열은 3글자 명사 하나로 이루어질 수도 있지만 2글자+1글자, 1글자+2글자 명사의 조합은 물론 1글자짜리 명사 3개의 조합으로 나타낼 수도 있다. 이 때, “기차길”, “기차”+“길”과 같이 조합하였을 때 동일한 결과가 나오는 원문은 중복으로 세지 않고 하나의 경우의 수로 계산하였고 그 결과는 표 4와 같다.

표 4에서 한글 초성 문자열의 길이가 증가할수록 복원 가능한 경우의 수가 기하급수적으로 증가하며 길이 k의

<표 2> 각 초성으로 구성된 한 글자 명사의 개수 분포

초성	ㄱ	ㄴ	ㄷ	ㄹ	ㅁ	ㅂ	ㅅ
개수	86	31	53	14	38	59	59
초성	ㅇ	ㅈ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
개수	75	57	29	21	28	28	47

<표 3> 각 초성으로 구성된 두 글자 명사의 개수 분포

	ㄱ	ㄴ	ㄷ	ㄹ	ㅁ	ㅂ	ㅅ	ㅇ	ㅈ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
ㄱ	405	96	194	227	229	281	439	438	410	232	2	86	117	223
ㄴ	107	18	55	59	59	82	121	114	115	59	6	29	35	54
ㄷ	199	52	76	88	85	125	167	145	160	95	11	36	57	69
ㄹ	7	5	10	7	8	18	9	9	8	3	10	12	9	1
ㅁ	195	44	105	83	100	118	214	165	178	93	8	39	53	95
ㅂ	278	57	123	134	116	147	273	243	253	145	15	78	73	140
ㅅ	387	87	173	184	186	235	359	354	317	191	20	110	115	189
ㅇ	475	106	228	238	248	277	504	461	470	237	25	105	120	248
ㅈ	379	75	141	175	170	212	354	358	314	192	15	65	105	196
ㅊ	241	38	86	87	93	121	210	191	182	91	9	38	42	119
ㅋ	14	11	21	28	13	23	23	18	10	7	5	22	19	4
ㅌ	98	19	26	40	29	49	94	74	81	31	10	14	24	46
ㅍ	91	17	61	59	56	65	129	92	95	52	14	34	22	46
ㅎ	229	44	121	111	113	153	270	239	204	94	11	45	62	115

<표 4> 길이 k인 한글 초성 문자열로부터 추측 가능한 명사 조합에 대한 경우의 수. 평균값이 45^k에 근접한다

k	1	2	3	4	5
최소	14	203	2,940	42,581	616,714
최대	86	7,428	651,561	55,412,028	4,785,971,788
평균	45	2,017	91,099	4,119,075	186,121,674
45^k의 값	45	2,025	91,125	4,100,625	184,528,125

한글 초성에 대하여 평균 약 45^k의 경우의 수를 검토하여야 한다는 것을 확인할 수 있다. 예를 들어 한글 200자 초성 문서로부터 원문 복원을 위해 검토해야 할 경우의 수는 평균 4.3×10^{330}가지이다. 일반 한글 문서의 크기가 최소 수백 자는 되는 점을 감안할 때 한글 초성 문자열만을 이용하여 원문을 복원하는 것이 용이하지 않음을 알 수 있다.

4. 검색 가능성에 관한 실험

한글 문서에서 초성만을 추출하여 구성한 문자열 내에서 기존의 방법들을 이용한 검색이 가능한 지에 대하여 실험적으로 증명하고자 한다. 먼저 초성으로만 이루어진 길이 k의 두 문자열 $X=<x_1,\cdots,x_k>$, $Y=<y_1,\cdots,y_k>$에 대하여 두 문자열 X, Y의 일치을 $S(X,Y)$를 식 3과 같이 정의하였다.

$$S(X,Y)=\frac{|\{i\,|\,x_i=y_i\}|}{k} \qquad (3)$$

예를 들어 길이 10의 두 초성 문자열 "ㄱㅊㅅㄴㅇㅇㅈㄱㄱㅇ"와 "ㅇㄹㅅㄴㅈㅇㄴㅇㄹㅇ"의 일치율은 표 5와 같이 총 10개의 문자 중 4개의 문자가 일치하므로 0.4이다.

<표 5> 두 문자열 간의 일치율 계산 방법 ($k=10$)

X	ㄱ	ㅊ	ㅅ	ㄴ	ㅇ	ㅇ	ㅈ	ㄱ	ㄱ	ㅇ	S(X,Y)
Y	ㅇ	ㄹ	ㅅ	ㄴ	ㅈ	ㅇ	ㄴ	ㅇ	ㄹ	ㅇ	
일치	×	×	○	○	×	○	×	×	×	○	4/10 = 0.4

검색 실험을 위하여 사용할 데이터는 표 6과 같다. 데이터는 21세기 세종 계획 원시 말뭉치[6]를 병합하여 만든 것으로 각각의 말뭉치에 대한 초성 문서를 생성한 뒤 해당 초성 문서상에서 초성으로 구성된 질의문을 이용하여 검색을 수행하였다. 질의문은 해당 말뭉치 내에 명백하게 존재하는 문자열과 다른 데이터에서 임의로 추출한 문자열을 길이 k별로 20개씩 추출하여 생성하여 말뭉치내의 모든 길이 k의 초성 부분문자열과의 일치율을 계산하였다.

<표 6> 실험에 사용된 데이터

말뭉치 #	크기	출처
1	한글 50,000,000자	21C 세종 계획[6]
2	한글 120,000,000자	

말뭉치 1에 대한 실험 결과는 표 7, 말뭉치 2에 대한 실험 결과는 표 8에 각각 나타내었다. 실험을 수행한 결과 두 말뭉치 데이터 모두 $k\geq15$일 때 대상 문서 내에 존재하는 문자열을 질의문으로 하여 검색한 경우 유일한 검색 결과를 얻을 수 있었다. 이는 적당한 길이 이상의 질의문이 주어진다면 해당 질의문에 해당하는 문자열이 탐색 대상 문서에 존재하는지 여부를 알아낼 수 있음을 의미한다.

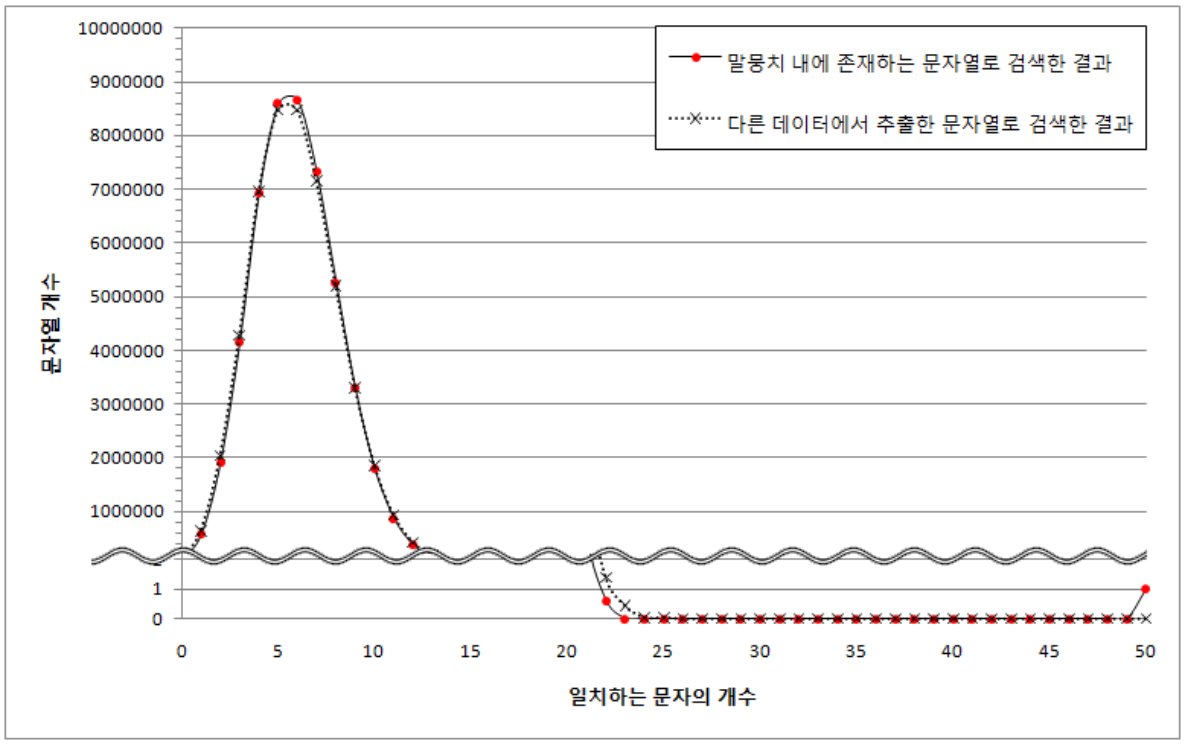
말뭉치 1 상에서 길이 50의 질의문으로 검색하였을 때 일치율에 따른 검색 결과 문자열 개수를 그림 1에 그래프로 나타내었다. 일치율에 따른 검색 결과 문자열 개수는

<표 7> 문자열 검색 결과 평균 개수 (말뭉치 1)

질의문	해당 말뭉치 내의 문자열					다른 데이터 내의 문자열				
일치율	0.0~0.2	0.2~0.4	0.4~0.6	0.6~0.8	0.8~1.0	0.0~0.2	0.2~0.4	0.4~0.6	0.6~0.8	0.8~1.0
k= 5	2.6×10^7	1.8×10^7	5.0×10^6	7.6×10^5	6.3×10^4	2.6×10^7	1.7×10^7	4.2×10^6	5.1×10^5	3.1×10^4
k=10	3.5×10^7	1.4×10^7	9.5×10^5	1.6×10^4	8.3×10^1	3.2×10^7	1.6×10^7	1.2×10^6	2.4×10^4	1.3×10^2
k=15	3.7×10^7	1.2×10^7	2.5×10^5	5.8×10^2	1.1×10^0	3.7×10^7	1.3×10^7	2.4×10^5	4.5×10^2	0.1×10^0
k=20	3.9×10^7	1.1×10^7	7.2×10^4	2.3×10^1	1.0×10^0	3.9×10^7	1.1×10^7	8.4×10^4	3.0×10^1	0.0×10^0
k=25	4.1×10^7	9.0×10^6	2.1×10^4	1.1×10^0	1.0×10^0	4.1×10^7	9.1×10^6	2.2×10^4	1.5×10^0	0.0×10^0
k=30	4.2×10^7	8.1×10^6	7.7×10^3	0.1×10^0	1.0×10^0	4.3×10^7	6.8×10^6	4.2×10^3	0.1×10^0	0.0×10^0
k=35	4.4×10^7	6.3×10^6	1.8×10^3	0.0×10^0	1.0×10^0	4.4×10^7	6.7×10^6	2.4×10^3	0.0×10^0	0.0×10^0
k=40	4.5×10^7	5.2×10^6	3.8×10^2	0.0×10^0	1.0×10^0	4.4×10^7	6.2×10^6	6.4×10^2	0.0×10^0	0.0×10^0
k=45	4.6×10^7	3.9×10^6	6.9×10^1	0.0×10^0	1.0×10^0	4.6×10^7	4.3×10^6	8.2×10^1	0.0×10^0	0.0×10^0
k=50	4.7×10^7	3.2×10^6	1.9×10^1	0.0×10^0	1.0×10^0	4.7×10^7	3.4×10^6	3.7×10^1	0.0×10^0	0.0×10^0

<표 8> 문자열 검색 결과 평균 개수 (말뭉치 2)

질의문	해당 말뭉치 내의 문자열					다른 데이터 내의 문자열				
일치율	0.0~0.2	0.2~0.4	0.4~0.6	0.6~0.8	0.8~1.0	0.0~0.2	0.2~0.4	0.4~0.6	0.6~0.8	0.8~1.0
k= 5	6.0×10^7	4.4×10^7	1.3×10^7	1.9×10^6	1.6×10^5	6.6×10^7	4.2×10^7	1.0×10^7	1.3×10^6	7.7×10^4
k=10	8.1×10^7	3.6×10^7	2.6×10^6	4.7×10^4	2.2×10^2	8.0×10^7	3.7×10^7	3.0×10^6	6.4×10^4	3.5×10^2
k=15	9.1×10^7	2.8×10^7	4.9×10^5	9.0×10^2	1.4×10^0	8.5×10^7	3.3×10^7	8.1×10^5	2.0×10^3	0.9×10^0
k=20	9.2×10^7	2.8×10^7	2.3×10^5	1.1×10^2	1.1×10^0	9.5×10^7	2.5×10^7	1.4×10^5	3.7×10^1	0.0×10^0
k=25	9.9×10^7	2.1×10^7	5.5×10^4	4.4×10^0	1.0×10^0	9.8×10^7	2.1×10^7	4.0×10^4	1.8×10^0	0.0×10^0
k=30	1.0×10^8	1.9×10^7	1.3×10^4	0.1×10^0	1.0×10^0	1.0×10^8	1.9×10^7	1.7×10^4	0.1×10^0	0.0×10^0
k=35	1.1×10^8	1.3×10^7	2.3×10^3	0.0×10^0	1.1×10^0	1.0×10^8	1.6×10^7	4.7×10^3	0.0×10^0	0.0×10^0
k=40	1.1×10^8	1.4×10^7	1.5×10^3	0.0×10^0	1.0×10^0	1.1×10^8	1.3×10^7	1.4×10^3	0.0×10^0	0.0×10^0
k=45	1.1×10^8	1.0×10^7	3.0×10^2	0.0×10^0	1.1×10^0	1.1×10^8	1.1×10^7	3.6×10^2	0.0×10^0	0.0×10^0
k=50	1.1×10^8	9.0×10^6	1.0×10^2	0.0×10^0	1.0×10^0	1.1×10^8	6.5×10^6	2.7×10^1	0.0×10^0	0.0×10^0



(그림 1) 말뭉치 1에서 길의 50의 초성 질의문으로 검색한 결과. 질의문의 검색 대상 말뭉치 내 존재 여부에 관계없이 비슷한 분포를 나타내며 완전 일치하는 문자열 개수에서 확실한 차이를 보인다.

질의문이 검색 대상 말뭉치 내에 존재하느냐 여부에 관계없이 비슷한 형태를 가졌으나 완전 일치하는 문자열의 존재 여부에만 차이가 확연히 드러났다. 그래프의 모양으로 미루어 볼 때 분포가 전체적으로 Poisson 분포를 따르므로 Poisson 분포의 파라메터 λ를 통해 질의문에 따른 검색 결과의 개수를 간단하게 유추할 수 있을 것으로 보인다.

5. 결론 및 검토

현대에 이르러 정보 검색의 중요성이 대두되는 동시에 정보 검색 시스템과 관련한 저작권 및 기타 사회 법률적인 분쟁이 문제되고 있다. 이 문제를 공학적으로 해결하기 위해서는 원문과 질의문의 노출을 방지하는 방법이 필요하다. 본 논문에서는 한글 문서로부터 초성만을 추출하여 나열한 초성 문서를 정의하고 이 방법에 의하여 변환된 문서가 원문 및 질의문의 노출을 방지할 수 있음을 명사 사전을 통한 단어 재구축 실험을 통하여 확인하였고, 실제 한글 말뭉치 상에서 검색을 수행하는 실험을 통하여 길이 15이상의 질의문에 대하여 유일한 검색 결과가 도출되는 것을 확인하여 검색에 충분히 활용될 수 있다는 것을 입증하였다.

다만 본 논문에서 사용한 일치율 계산 방법에 따르면 질의문과 검색 대상 간의 단순 음절 불일치만 있는 경우가 아니라 삽입 또는 삭제 부분이 존재하는 경우 유사한 문자열임에도 불구하고 검색에 실패할 수 있다. 이 경우 질의문을 적당한 길이로 절단하여 검색하는 방법을 사용할 수 있는데 일치율에 따른 검색 결과 분포가 Poisson 분포를 따르므로 파라메터 λ를 통하여 적합한 절단 길이를 예측할 수 있을 것으로 보인다.

Acknowledge

본 연구는 2010년도 연구재단 일반연구자지원사업 (2010-0015665)의 연구비 지원으로 수행하였습니다.

참고문헌

[1] Christopher D. Manning, Prabhaker Raghaven and Hinrich Schütze, "Introduction to Information Retrieval," Cambridge University Press, 2008.

[2] 김윤명, "정보검색 서비스에 관한 저작권법적 고찰," 산업재산권, vol.24, pp.269-305, 2007.

[3] William Gasarch, "A survey on private information retrieval," Bulletin of the EATCS, vol.82, pp.72-107, 2004.

[4] 김선영, 이필중, 서재우, "검색 가능 암호 기술의 연구 동향," 정보보호학회지, vol.19, no.2, pp.63-73, 2009.

[5] 이재홍, 오상현, "한글 음절의 초성, 중성, 종성 단위의 발생확률, 엔트로피 및 평균상호정보량," 전자공학회논문지, vol.27, no.9, pp.1299-1307, 1989.

[6] 21세기 세종계획, http://www.sejong.or.kr/