음성데이터베이스로부터의 효율적인 색인데이터베이스 구축과 정보검색

The Extraction of Effective Index Database from Voice Database and Information Retrieval

박 미 성(Mi-Sung Park)*

<목 차>

초 록

전자도서관과 같은 정보제공원은 이미지, 음성, 동영상 등과 같은 비정형 멀티미디어 데이터 서비스에 대한 요구를 받고 있다. 그러하여 본 연구에서는 음성 처리를 위해 어절생성기, 음절복원기, 형태소분석기, 교정기를 제안하였다. 제안한 음성처리 기술로 음성데이터베이스를 텍스트데이터베이스로 변환 한후 텍스트데이터베이스로부터 색인데이터베이스를 추출하였다. 그리고 추출한 색인데이터베이스로 텍스트와 음성의 내용기반정보검색에 활용할 수 있음을 보이기 위해 정보검색모델을 제안하였다.

주제어 : 음성데이터베이스, 색인데이터베이스, 어절생성기, 음절복원기, 색인추출, 정보검색

Abstract

Such information services source like digital library has been asked information services of atypical multimedia database like image, voice, VOD/AOD. Examined in this study are suggestions such as word-phrase generator, syllable recoverer, morphological analyzer, corrector for voice processing. Suggested voice processing technique transform voice database into text database, then extract index database from text database. On top of this, the study suggest a information retrieval model to use in extracted index database, voice full-text information retrieval.

Key Words : voice database, index database, word-phrase generator, syllable recoveror, index extraction, information retrieval

* 경북대학교 중앙도서관 전산관리팀 팀장(mspark@kmu.ac.kr)
·접수일 : 2004. 9. 1 ·최초심사일 : 2004. 9. 4 ·최종심사일 : 2004. 9. 12

Ⅰ. 서 론

소용돌이치는 21세기는 바야흐로 정보화 사회의 정보 기반 구조로서, 고속 정보망의 구축, 개인용 컴퓨터의 급속한 보급, 멀티미디어 기술의 발전 등으로 인하여 과거에는 상상할 수 없었던 정보서비스의 새로운 장을 열고 있다. 오늘 날 엄청나게 많은 전자화된 정보 자원들이 웹을 통해 유통과 재생산을 반복하고 있고, 이에 따라 데이터베이스 연구자들은 다양한 방법으로 가상(Cyber)공간에 존재하는 방대한 디지털 웹 자원을 대상으로 조직적 관리, 체계화 및 표준화를 꾀하고 있다. 이러한 방대한 디지털 웹 자원에는 지금까지 주요 접근 대상이었던 문서만이 아니라 이미지, 음성, 동화상과 같은 다양한 멀티미디어 정보를 포함하고 있다 따라서 다양한 이용자 서비스를 위해서는 다양한 멀티미디어 자료의 처리 및 저장, 분류, 그리고 검색 등의 분야에 많은 연구가 이루어져야 한다.[1)]

일반적으로 다양한 정보자원의 공급원을 생각하면 도서관을 떠올리게 되는데, 이러한 도서관은 오늘날과 같이 급변하는 정보기반사회를 맞이하면서 정보자원의 공급 기관으로서의 서비스에 대한 많은 숙제를 안게 되었다. 수작업으로 처리하던 도서관 관리업무는 약 10년 전을 깃점으로 자동화를 도입하기 시작하여 그 이후 서지정보에 대하여 마크(KORMARC, USMARC 등) 형식의 데이터베이스를 구축하면서 도서관 이용자들을 위하여 간단한 서지검색 및 CD-ROM 검색 정도의 서비스를 제공하여 왔으나, 지금은 단순한 서지검색 차원을 넘어서 원문, VOD, AOD, 이미지, 음성, 동영상, 전자저널, 웹 자원 등 다양한 형식의 전자자원에 대하여 서비스를 해야만 하는 입장에 처하게 되었다. 이처럼 정형화된 데이터뿐만이 아니라 다양한 비정형 멀티미디어 자원을 구축하고 이를 서비스해야 함으로 인하여, 오늘날 도서관의 개념은 전자도서관, 디지털 도서관, 가상도서관으로 전환되고 있다.

최근 이러한 전자도서관은 그 서비스의 특성에 따라 3가지 유형으로 세대 구분이 이루어지고 있다.[2)]

1세대 전자도서관 시스템은 이미지, 동영상, 웹자원, 원문(Full-text) 등 다양한 형식의 전자 자원을 효율적으로 관리하는 시스템으로서 전통적인 도서관 자동화 시스템의 부가 또는 확장 기능에 초점을 두고 있으며, 또한 TEI, EAD, CDWA, Dublin Core 등 여러 가지 디지털 자원의 정보 표현을 위한 메타데이터에 대한 표준 등장에 그 특징이 있다.

2세대 전자도서관은 1세대 보다 약간 진화된 전자도서관 모델로, 조직체 내에 산재해 있는 다양한 디지털 자원의 통합 관리 및 포탈 기반의 통합서비스 및 개인화 서비스 요구를 수용하는 시스템을 구축하고자 한다는 점이다. 이 세대는 다양한 유형의 메타데이터 통합을 위한 세대로

1) 홍기형, 박치항, "전자 도서관의 요소 기술 : DBMS와 정보검색," 정보과학회지, 제15권, 제2호(1997, 2), p.13.
2) 이수상, "전자도서관의 최근동향," 데이터베이스 연구, 18권, 3호 (2002, 9), pp.4-5.

현재 대부분의 국내외 전자도서관이 이 경우에 해당된다고 하겠다.

3세대 전자도서관은 현재 실험적으로 구축 중인 개념으로 분산환경의 전자 도서관들간 통합 및 연동성 문제를 강조하고, 전자도서관 상호간의 개방성, 표준성, 상호운용성에 관한 내용들을 핵심 이슈로 부각시키고 있다.

이처럼 전자도서관이 서비스 유형별로 세대 구분이 이루어지고 있는 이 즈음, 시간이 갈수록 전자도서관은 다양한 자원 즉 이미지, 음성, 동영상과 같은 방대한 멀티미디어 정보 처리에 대한 요구를 받고 있다. 따라서 성공적인 전자도서관 구축을 위하여 멀티미디어 데이터에 대한 체계적이고 조직적인 데이터베이스 구축과 함께 다양한 자료 제공 방법에 대한 정보 검색 기술이 필수적인 요소로 대두되고 있다.

그러므로 본 연구에서는 멀티미디어 정보 중 대용량 음성데이터베이스를 문자데이터베이스로 변환 구축하고, 변환된 대용량 문자데이터베이스를 대상으로 색인데이터베이스를 구축한 후, 정보 검색에서 이용자의 음성 및 문자를 대상으로 내용 기반 검색이 가능하게 하는 기반 모델에 대해 제안하고자 한다.

이에 따른 본 연구의 내용은 다음과 같다. Ⅱ장에서는 음성 데이터베이스의 문자데이터베이스 변환기술에 관련된 정보와 처리기술에 대해 소개하고, Ⅲ장에서는 음성데이터베이스에서 추출된 문자데이터베이스를 대상으로 효율적인 색인데이터베이스 구축 기법에 대해 소개하며, Ⅳ장에서는 Ⅱ장과 Ⅲ장에서 제안한 기반기술이 적용되어 정보 검색에 활용될 수 있는 정보검색 모델에 대해 논의한 후, 마지막 Ⅴ장에서 결론을 맺는다.

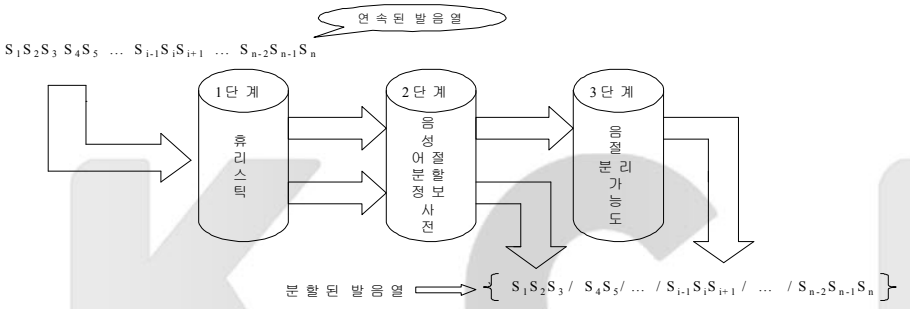
Ⅱ. 음성데이터베이스의 텍스트데이터베이스로의 변환

사람이 발화한 음성데이터베이스를 처리하여 문자데이터베이스로 변환하고자 할 때 일반적으로 발생하는 몇 가지 문제점이 있는데, 그 중 한가지는 자연스럽게 발화되는 연속음성을 적절한 구획으로 분할해야 하는 문제이고, 다른 한가지는 한국어의 특성상 연속해서 발화할 때 형태소간에 음운변동이 일어나 발음열과 자소열이 다르게 나타난다는 것이다. 따라서 본 장에서는 위의 첫 번째 문제해결을 위한 어절생성기와 두 번째 문제해결을 위한 음절복원기의 전반적인 처리과정을 보이고, 어절생성기와 음절복원기에서 생성된 결과를 형태소 분석과 교정기를 거침으로 음성데이터베이스가 텍스트데이터베이스로 완성되어 나가는 과정에 관하여 살펴보고자 한다.[3)]

3) 박미성, 한국어 STT를 위한 연속 발음열 분할 및 자소열 복원시스템(박사학위논문, 경북대학교 대학원 컴퓨터공학과, 1999), p.101.

1. 어절생성기

어절 생성기는 발음열 분할을 위해 세 단계의 분리 위치 추정 단계를 거치도록 설계하였다. 첫 단계는 휴리스틱 정보를 이용한 1차 분리위치 추정단계로서 연속된 발음열에서 반드시 분할되어야 할 위치를 찾아내는 과정이다. 둘째 단계는 음성 어절 분할 정보 사전을 이용한2 차 분리위치 추정단계로서 최장 조사·어미 선택 원칙에 의거하여 분리위치를 결정한다. 셋째 단계는 한 음절을 사이에 두고 분리위치가 추정되었을 경우에 두 분리위치 중 좀 더 정확한 분리위치를 결정해 주는 단계이다. 세 번째 단계를 수행하는 이유는, 보통 한 어절이 조사·어미로 추정되어 한 음절로 구성될 가능성은 희박하기 때문에 이 단계를 수행하므로 한 음절씩 여러 개 분리된 경우에 이들을 결합해 주기 위함이다. 연속된 음성데이터베이스를 입력으로 받아 세 단계의 과정을 거쳐 어절에 근접한 수준으로 분할해 나가는 어절 생성기의 흐름도는 <그림 1>과 같다.



<그림 1> 어절 생성기의 흐름도

다음은 어절생성기의 각 단계에서 사용될 정보에 대하여 살펴본다.

1) 휴리스틱 정보

『휴리스틱』이란 주어진 문제의 해를 결정적인 알고리즘에 의하지 않고 시행착오를 통해 축적된 경험적인 정보이다.[4] 대개 인공지능 분야의 문제 해결을 위해 많이 사용되며, 각 분야별로 구해지는 휴리스틱의 종류도 다양하다. 본 연구에서는 문서교정시스템이나 자동띄어쓰기시스템[5]에 사용되는 휴리스틱을 연속된 발음열에서 어떻게 나타나는 가를 조사한 후, 각 음성 휴리스틱마다 앞 음절의 종성에 따라 어떠한 음운 변동이 일어나는 지를 예측하여 발생 가능한 모든 조

4) 컴퓨터 용어 사전 편찬위원회, 컴퓨터 용어 대사전(서울 : 크라운출판사, 1991).
5) 최재혁, "양방향 최장일치법을 이용한 한국어 띄어쓰기 자동 교정 시스템," 한글 및 한국어 정보처리 학술대회 발표논문집, 제9회(1997, 10), pp.145-151.

건을 고려하여 정보를 추출하였다. 이 정보를 『음성 휴리스틱 정보』라 하고 연속된 발음열을 대상으로 1차 분할위치를 추정해 내는데 이용하였다. 추출한 음성 휴리스틱 정보의 일부 내용을 살펴보면 <표 1>과 같다.

<표 1> 음성 휴리스틱 정보

▶ H_1 : -[기/끼/키]+ 때무네 …
▶ H_2 : -[기/끼/키]+ 위한 …
▶ H_3 : -ㄴ+[거/걷] …
▶ H_4 : -ㄴ+[데]
▶ H_5 : -ㄴ+[뒤]
▶ H_6 : -ㄴ+[지]
▶ H_7 : -ㄴ+[채]
▶ H_8 : -ㄴ+[휘]
▶ H_9 : -는+
▶ H_{10} : -ㄹ+[거/꺼/껃]
▶ H_{11} : -ㄹ+[떼]
▶ H_{12} : -ㄹ+[쭈/쭐]
▶ H_{13} : -ㄹ+[찌]
▶ H_{14} : -ㄹ+쑤+[이/인/일/엄/업-]
▶ H_{15} : -를+
▶ H_{16} : -에[게,네,레,메,베・・]+대한[대하여-, 대해서-] ▶ H_{17} : -에[게,네,레,메,베・・]+의한[의하여-, 의해-]
(+ : 분리 위치)

2) 음성 어절 분할 정보 사전

음성 어절 분할 정보 사전은 연속 발음열의 2차 분리위치 추정정보를 담고 있는 사전으로, 한국어 어절구성의 특징과 음운변동을 근거로 하여 만든 사전이다. 한국어 어절구성은 보통 '체언+조사'나 '용언+어미'로 결합되어 있는 경우가 대부분을 차지하고, 관형사, 부사 등은 주로 단독어절을 형성하는 특성을 갖는다. 그러므로 어절을 구성하고 있는 형태 중 뒷부분인 『조사』나 『어미』 그리고 단독어절을 형성하고 있는 『관형사』, 『부사』의 음절은 어절의 끝을 추정하기 위한 정보로 충분히 이용될 수 있다. 이와 같은 사실을 하나의 근거로 삼고 사전의 표제어로 조사, 어미, 관형사, 부사의 발음열을 등재하였다. 그리고 사전구성을 위한 또 다른 근거가 되는 특성은 음운변동이다. 보통 한국어는 발음할 때 음소결합의 제약성, 발음의 편의, 말의 청취효과에 따른 명확성 등의 이유로, 한 음절의 초성과 중성, 중성과 종성 그리고 종성과 다음 음절의 초성 사이에 음운 변동이 일어난다는 사실이다. 음운변동에 대한 간단한 예를 살펴보면 <표 2> 와 같다.

<표 2> 앞 음절 종성과 다음 음절 초성 사이의 음운변동

밥을 먹고 ---> 바블 먹**꼬** : ㄱ + 고 ---> ㄱ + **꼬**
공부하고 ---> 공부하 **고** : # + 고 ---> # + 고
놀지 않고 ---> 놀지 안**코** : ㄶ + 고 ---> ㄴ + **코**

따라서 어절의 분할 정보로 사용하기에 충분한 이러한 두 가지 특징들을 고려하여 조사・어미, 부사, 관형사 등의 음절에 대해 각각의 음운변동 조건과 그에 따라 변동된 음운이 수록된 6000개의 음성 어절 분할 정보 사전을 구축하게 되었다. 구축한 음성 어절 분할 정보 사전의 일부 예를

보이면 <표 3>와 같다.

<표 3> 음성 어절 분할 정보 사전

연속된 발음열	앞 음절의 종성 조건
가	#
고	#,ㄹ
고는	#,ㄹ
고자	#,ㄹ
과	dㄴ,ㄹ,ㅁ,ㅇ
:	:
기에	#,ㄹ
긴	#
께서	#,ㄱ,ㄴ,ㄷ,ㄹ,ㅁ,ㅂ,ㅇ
:	:

2차 분할위치 결정을 위해 좌 방향에서 우 방향으로 표제어의 발음열을 검색하여 입력된 발음열과 매칭하되, 앞 음절 종성 조건이 만족되는 범위 내에서 가장 긴 조사·어미 발음열을 선택하게 하는 최장 조사·어미 선택 정책을 사용한다. 음성 어절 분할 정보 사전을 참조하여 분할이 수행되는 과정은 <그림 2>와 같다.

```
for(i=0; i<buffer.length; i++)
        current <- buffer[i]
        before <- buffer[i-1];
        next <- buffer[i+1];
        coda <- before.coda; // 앞 음절의 종성
        if(current != Heuristic) //휴리스틱에서 분리한 정보가 아니면
                for(j=5; j>0; j--)
                        for(k=0; k<j ; k++)
                                temp[k] <- buffer[k]
                        if(isExistDictionary(temp))
                                makeSpace(buffer,i) ;
```

<그림 2> 음성 어절 분할 정보 사전 참조 알고리즘

3) 음절 분리 가능도

음절 분리 가능도는 발음열의 3차 분리위치 추정정보로 사용되는데, 2차 단계까지 분할된 위치 중에서 한 음절을 사이에 두고 분리 위치가 추정된 경우에 올바른 위치를 잡아 주기 위해 이용되는 정보이다. 이 정보는 한국어의 조사·어미 음절이 특성상 때때로 체언이나 용언의 일부가 될 수 있는데, 이처럼 체언, 용언의 일부가 되는 조사·어미로 인식되어 분리 위치가 잘못 추정되거나 두 군데 이상의 분리점이 생겨 모호성이 발생한 경우에 대한 해결책으로 사용된다. 제안한 음절 분리 가능도 $P(Si)$는 "음절 Si가 끝음절로 사용될 가능성"이라 정의하였고, 다음의 식

(1)에 의해 값을 구하였다.

$$(1)\ \text{음절 분리 가능도} \quad P(Si) = \log \frac{Pe(Si)}{Pf(Si)}$$

식(1)에서, $Pe(Si)$는 "Si가 어절의 끝음절로 사용될 확률"이라 정의하였고, 약 60만 어절의 말뭉치에서 추출하였으며 $Pe(Si)$는 다음 식(2)에 의해 계산된다.

$$(2)\quad Pe(Si) = \frac{Sifi}{S1f1 + S2f2 + \cdots + Sifi + \cdots + Snfn}$$

식(2)에서, $S1, S2, Si, Sn$은 말뭉치에서 끝음절로 사용되는 각각의 음절을 나타내고, $f1, f2, fi, fn$은 말뭉치에서 끝음절로 사용된 각 음절의 빈도값이다. 그리고 $Pf(Si)$는 Si가 체언이나 용언의 첫 음절로 사용될 확률을 나타내는 것으로 약 23만 단어가 수록된 어휘사전에서 모든 어휘를 음성 발음열로 변환시킨 뒤 추출하였다. $Pf(Si)$는 식 (3)과 같이 계산된다.

$$(3)\ Pf(Si) = \frac{Sifi'}{S1f1' + S2f2' + \cdots + Sifi' + \cdots + Snfn'}$$

식(3)에서, $f1', f2', fi', fn'$는 어휘사전에서 첫 음절로 사용된 각 음절의 빈도값이다. 이렇게 구한 음절 분리 가능도의 개수는 약 1,527개인데 일부를 보이면 <표 4>와 같다.

<표 4> 음절 분리 가능도

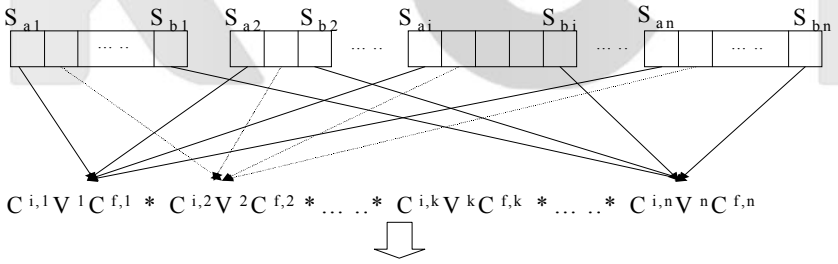
순위	음절	$Pe(Si)$	$Pf(Si)$	$P(Si)$
1	를	6.997	0.001	3.844912
2	른	1.584	--	3.199755
3	는	9.195	0.006	3.185259
4	떤	0.912	--	2.959995
5	믈	1.683	0.002	2.924796
6	릐	0.717	--	2.855519
132	에	2.191	0.529	0.617210
140	한	1.415	0.394	0.555215
164	가	3.277	1.222	0.465531
357	다	0.366	0.830	-0.355561
487	자	0.168	1.100	-0.815309
500	상	0.088	0.676	-0.886056

다음 <그림 3>은 음절 분리 가능도를 이용하여 분할된 대상을 결합해 주는 알고리즘이다.

```
for(i=0; i<buffer.length; i++)
        current <- buffer[i]
        before <- buffer[i-1];
        next <- buffer[i+1];
        coda <- before.coda; // 앞 음절의 종성
        if(before='#' and next='#') // 앞과 뒤가 모두 띄어쓰기 할 대상이면
                if(before.size > next.size)
                        frontMakeSpace(buffer,i);
                else if(before.size < next.size)
                        rearMakeSpace(buffer,i);
                if(current.size=next.size=1)
                                // 음절 분리 가능도 비교
                        if(Frequency(current) > Frequency(next))
                                frontMakeSpace(buffer,i);
                        else
                                rearMakeSpace(buffer,i);
```

<그림 3> 음절 분리 가능도 참조 알고리즘

어절생성기는 이상과 같은 세 단계의 처리 과정을 거쳐 연속된 음성 발음열을 어절이라 추정되는 적절한 크기로 분할시켜 준다. 분할위치가 결정되면 다음 단계의 음절 복원기에서는 분할된 음성열을 기본 입력으로 받아 복원을 수행한다. 복원기의 입력은 다음의 <그림 4>의 식(4)와 같이 표현할 수 있다.



(4) $S_a \cdots S_b = C^{i,1}V^1C^{f,1} * C^{i,2}V^2C^{f,2} * \cdots * C^{i,k}V^kC^{f,k} * \cdots * C^{i,n}V^nC^{f,n}$

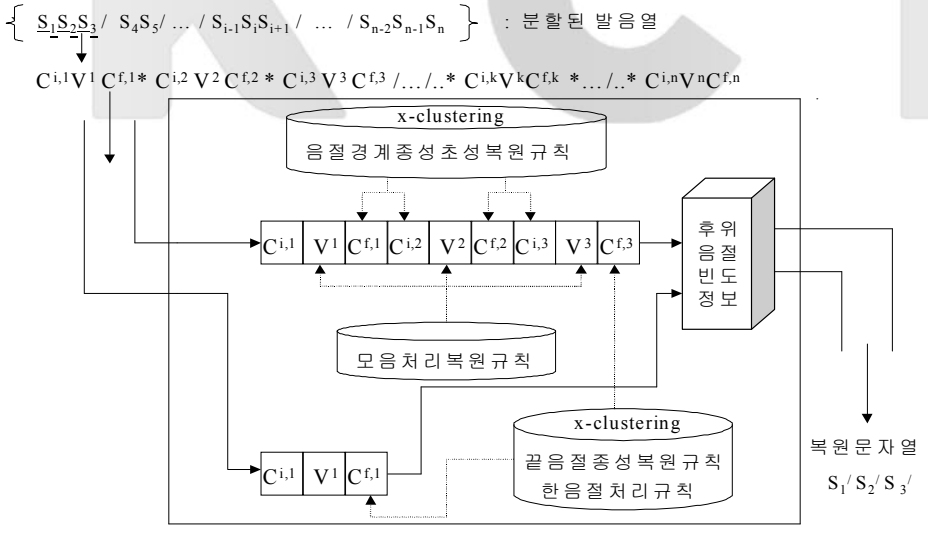
<그림 4> 추정 어절의 음절 복원기 입력 패턴으로의 대응 관계

<그림 4>에서 연속 음성 발음열 $S_1S_2 \cdots\cdots S_{n-1}S_n$는 어절생성기를 통해, 몇 개의 $S_a \cdots S_b$(1 <= a, b < n, a<=b) 크기로 분할된다. 분할된 각 어절 $S_a \cdots S_b$는 음절복원기의 입력으로 사용되기 위해

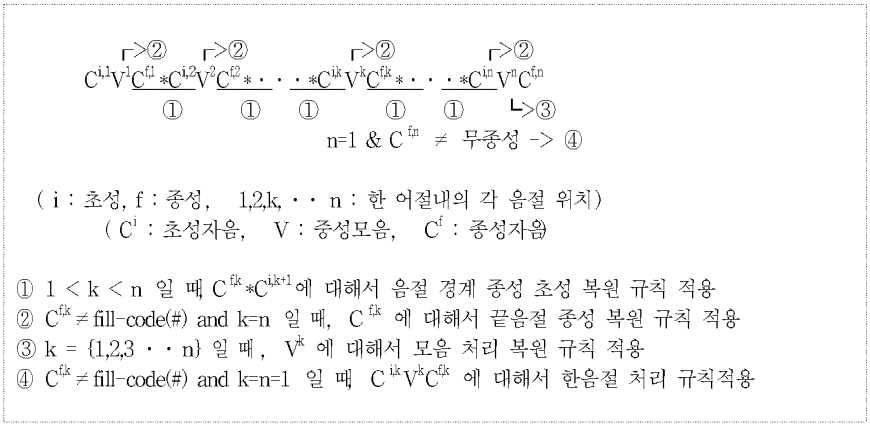
서 입력 패턴이 초성(C^i)중성(V^i)종성(C^f)의 형태로 변환된다. 복원기의 입력은 어절 단위이므로 어절생성기의 S_a가 항상 첫 번째 음절로, S_b는 n번째 음절로 대응되어 $S_a=C^{i,1}V^1C^{f,1}$, $S_{a+1}=C^{i,2}V^2C^{f,2}\cdots$, $S_b= C^{i,n}V^nC^{f,n}$ 로 표현된다. 식(4)에서 i는 초성, f는 종성을 의미하며, 1, 2, k, n은 한 어절내의 각 음절 위치 즉 첫 번째, 두 번째, k 번째, n 번째 음절을 의미한다. 그러므로 $C^{i,1}$은 첫 음절 초성을, V^1은 첫 음절 중성을, $C^{f,1}$은 첫 음절 종성을 가리킨다.

2. 음절복원기

음절 복원기는 어절 생성기에서 생성한, 예측된 어절들을 입력으로 받아 정해진 음절 복원 rule-base에 의거하여 복원을 수행한다. 이 때 좀 더 효과적인 복원을 위해 x-clustering정보와 후위(postfix) 음절 빈도 정보를 참조한다. 음절 복원기의 상세 흐름도는 <그림 5>와 같다. <그림 5>에서 음절 복원기의 입력은 앞의 식(4)의 형태 즉 분할된 어절의 음소열 형태로 입력이 된다. 입력된 음소열의 각 위치와 조건에 따라 음절 경계 종성 초성 복원 규칙, 모음 처리 복원 규칙, 끝음절 종성 복원 규칙, 한 음절 처리 규칙이 적용되며 종성 복원이 이루어질 때는 x-clustering 정보 참조가 이루어져 올바른 음절을 생성할 수 있고, 복원이 완료된 후에는 후위 음절 빈도 정보 참조로 복원 후보 수를 줄여주는 과정으로 진행된다. 복원 규칙이 적용되는 상세한 조건은 <그림 6>과 같다.



<그림 5> 음절 복원기의 흐름도



<그림 6> 복원 규칙이 적용되는 상세 조건

위 <그림 6>에서 적용되는 전체 복원규칙[6]은 91개의 주 규칙과 240개의 부 규칙으로 이루어진다.

1) 음절 경계 종성 초성 복원 규칙

앞 음절 종성과 다음 음절의 초성 사이에 적용될 수 있는 규칙으로, 음성 발음열에서 종성으로 나타날 수 있는 것은 무종성(fill-code 또는 #)과 7개의 대표음 'ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅇ' 이므로 이를 기준으로 규칙을 정의하였다. 여기에 해당하는 규칙은 83개의 주 규칙(main rule) 과 227 개의 부 규칙(sub path)이 있다.

2) 모음처리 복원 규칙

모음 처리 복원 규칙은 모음의 복원을 위해 사용되는 규칙으로 크게 3개의 주 규칙(main rule) 을 가진다.

3) 끝음절 종성 복원 규칙

한 어절의 마지막 음절에 종성이 있을 경우에 적용되는 규칙으로, 5개의 주 규칙(main rule)과 13개의 부 규칙(sub path)으로 구성된다.

6) 박미성, 전게논문, pp.41-51.

4) x-clustering 정보

상용조합 2,350자를 대상으로 같은 종성을 가지는 무리끼리 그룹을 만들어 x 받침을 가질 수 있는 초성, 중성 쌍들의 집합을 x-clustering 정보라 정의하였다. 이 정보는 이 정보를 이용하면 복원 시 올바른 음절만을 생성시켜 불필요한 복원 과정(path)은 거치지 않도록 미연에 방지하는 효과를 가져올 수 있었다.

5) 후위 음절 빈도 정보

규칙에 의해 복원된 후보 수는 보통 한 개부터 여러 개까지 나타났다. 형태소 분석 전에 나온 후보들을 조사해 보았더니 한 음절 한 음절은 문자로 만들어질 수 있지만 그 음절들이 모여서는 의미를 갖지 못하는 후보들이 많았다. 복원 후보가 많이 생성된 경우에 단어로 인정될 수 없는 후보까지도 형태소 분석을 한다면 형태소 분석기의 처리 시간이 증가하게 되어 시스템의 전체 수행 시간에 영향을 미치게 된다. 그러므로 형태소분석기의 입력 후보를 제한하여 단어로 인정될 수 없는 후보를 미리 필터링하기 위한 방안으로 후위 음절 빈도 정보를 정의하여 사용한다. 임의 음절에 대한 후위 음절 빈도 정보 값은 그 음절 뒤에 바로 나타날 수 있는 문자들의 빈도 값으로 60만 말뭉치로부터 구했고, 이 정보 값을 이용해 특정 음절 뒤에 올 수 있는 음절과 없는 음절을 판단하여 많은 후보를 줄일 수 있었다.

3. 형태소분석기

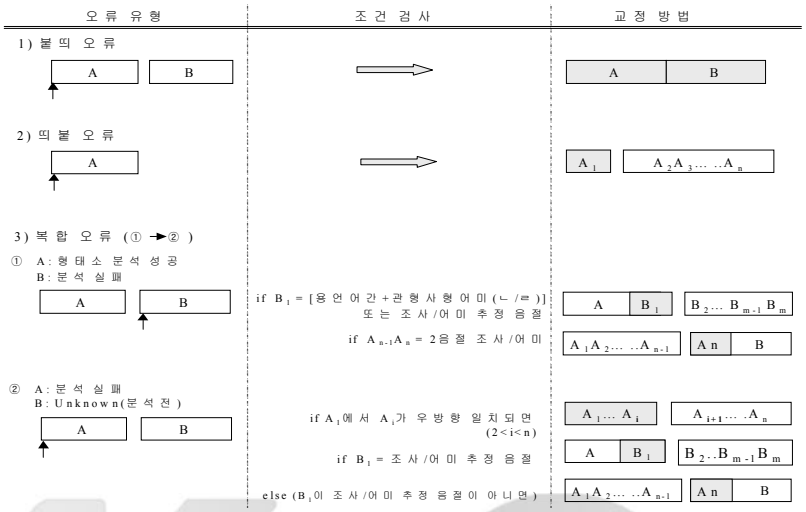
어절생성기와 음절복원기의 처리 과정을 통해 생성된 대용량 텍스트데이터베이스에 대해 문장 단위로 형태소분석을 하게 된다. 본 연구에서는 형태소분석을 위해 O(n)의 시간 복잡도를 가지는 Top-down 방식의 양방향 최장일치 형태소분석기[7]를 이용하여 형태소 분석을 한 후 다음 교정 단계에서 교정을 수행하도록 하였다.

4. 교정기

분할된 어절은 규칙에 의거하여 복원된 후 형태소 분석되는데, 만약 어떤 어절에 대하여 형태소 분석에 성공한 후보가 하나도 존재하지 않을 때 본 시스템은 어절 분할이 잘못된 것으로 간

7) 최재혁, 양방향 최장일치법에 의한 한국어 형태소 분석기의 구현(박사학위논문, 경북대학교 대학원 컴퓨터공학과, 1993), p.94.

주하고 적절한 교정을 수행하게 되는데 이 역할을 시스템의 교정기가 담당하게 된다. 교정의 대상이 되는 오류는 주로 어절의 경계를 잘못 추정하여 생기는 어절분할오류이다. 본 연구에서는 이러한 어절분할오류를 오류의 유형에 따라 붙띄오류, 띄붙오류, 복합오류 세 가지로 분류하였다.[8] 세 가지 오류의 유형별 교정방법을 살펴보면 <그림 7>과 같다.



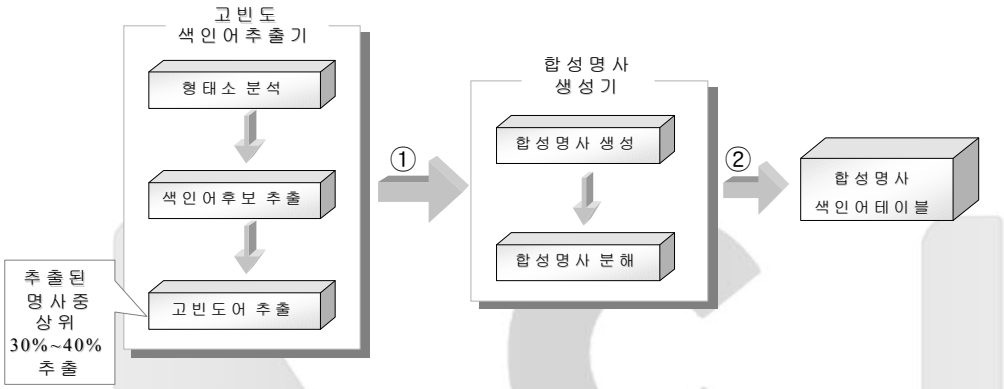
<그림 7> 오류 분석 및 교정 방법

먼저 교정 시 교정 대상이 되는 두 어절을 A, B라 하고 각각 m, n개의 음절로 구성되어 있다고 할 때, A는 $(A_1A_2...A_{n-1},A_n)$로, B는 $(B_1B_2...B_{m-1},B_m)$로 표현한다. 그리고 ↑는 현재 분석 위치를 나타낸다. <그림 7>에서 붙띄오류는 A, B중 하나 또는 둘 다 형태소 분석에 실패하고, 두 어절길이의 합이 6어절 이하(n+m<=6) 일 때 두 어절을 결합시켜 교정한다. 띄붙오류는 여러 어절이 한 어절로 결합된 경우로 한 음절씩 잘라서 적절한 분할 위치를 찾는다. 복합오류는 조건에 따라 앞 어절의 맨 마지막 음절을 잘라서 뒤 어절에 결합시키거나, 뒤 어절의 첫 음절을 잘라서 앞 어절의 뒤에 붙이는 식으로 교정이 이루어진다. 이 과정이 완료되면 음성데이터베이스의 변환된 텍스트데이터베이스가 생성된다.

8) 박미성, 전게논문, pp.60-62.

Ⅲ. 텍스트데이터베이스에서 색인데이터베이스 구축

대량의 웹 문서를 검색하는 시스템이나 향후 전자도서관의 음성기반 자연어 인터페이스 시스템 구현과 같은 음성 기반 검색에 있어서 추출된 색인어의 정확도가 시스템 성능 평가에 중요한 위치를 차지하고 있다. 따라서 이 장에서는 앞장의 대용량의 음성데이터베이스를 대상으로 생성한 텍스트데이터베이스를 입력으로 받아 고빈도어를 이용한 색인데이터베이스[9]를 구축함으로 사용자의 음성 질의 또는 문자 질의에 대한 효율적인 검색이 가능한 기반 기술을 제안하고자 한다. 구성도는 <그림 8>과 같다.



<그림 8> 합성 명사 생성의 구성도

정보검색에서 합성명사를 적절하게 처리하는 것은 검색시스템의 효율을 향상시키는데 매우 중요하다.[10] 이 말은 문서 내에서 명사가 차지하는 개념적 중요도가 크고, 대부분 색인어로 사용되고 있음을 말해준다. 본 연구에서는 효율적인 색인어를 추출하기 위해 상위 30%~40%의 출현 고빈도어를 대상으로 다양한 합성규칙과 분해규칙을 적용하여 합성명사 색인데이터베이스를 생성한다.

9) 김미진, 박미성, 최재혁, 이상조, "효율적인 색인어 추출을 위한 합성명사 생성 방안에 대한 연구," 한국정보처리학회, 제7권, 제4호(2000. 4), pp1123-1127.
10) 신동욱, "복합명사의 통계적 처리에 대한 평가," 한글 및 한국어 정보 처리 학술발표논문집(1997. 10), pp.36-41.

1. 고빈도 색인어 추출기

대량의 변환된 텍스트데이터베이스를 입력으로 받아 간단한 형태소 분석을 통해 명사를 추출한다. 추출된 명사로부터 색인어 후보가 될 수 있는 모든 단일 명사와 복합 명사를 추출해 낸다. 추출한 후보 중 불용어는 불용어 사전을 통해 제거된다. 기존의 색인어 추출 시스템은 문서 내에 출현한 모든 단일 명사를 합성 명사 구성 조건에 의해 복합 명사 색인어로 추출함으로 너무 많은 색인어가 추출되어 시스템의 정확률을 저하시키는 원인이 된다. 그러므로 본 연구에서는 추출된 후보들 중에서 출현 빈도가 높은 상위 30%~40%에 대해서만 명사 합성을 수행한다. 고빈도어 추출을 위한 빈도 계산은 식(6)에 의해 이루어지고 고빈도어 추출 알고리즘은 <그림 9>와 같다.

$$(6)\ \mathrm{Termfreq}_{ij} = \frac{\text{문서}i\text{에서의 명사}j\text{의 출현횟수}}{\text{문서}i\text{에서 추출된 전체 명사개수}} * 100$$

```
[알고리즘 1] High_freq_extraction
Input    형태소 해석결과        A1, ... , An
         원하는 빈도            P1, ... , Pn
Output   고빈도어 추출결과      H1, ...,    Hn

High_freq_extraction_mehod
    for i = 0 to n do
     Hi.count = Ai.count;
     for j = i+1 to n do
       if (Aj.pumsa == 'N')
          if (Aj.data == Ai.data)
             Hi.data = Ai.data;
             Hi.count = Ai.count + Hi.count;
          end if
       end if
    enddo
  enddo
```

```
Q_Sort(H);
 for i = P1 to Pn do
   j = 0, bound = 0;
      do {
      bound = 고빈도_bound(H);
      //빈도의 누적값을 보여준다.
            j++;
      } while( bound < Pi);
      if(((Pi)-bound>bound 다음값-(Pi))
         고빈도_출력(H, bound);
      else
         고빈도_출력(H, bound);
      end if
 enddo
```

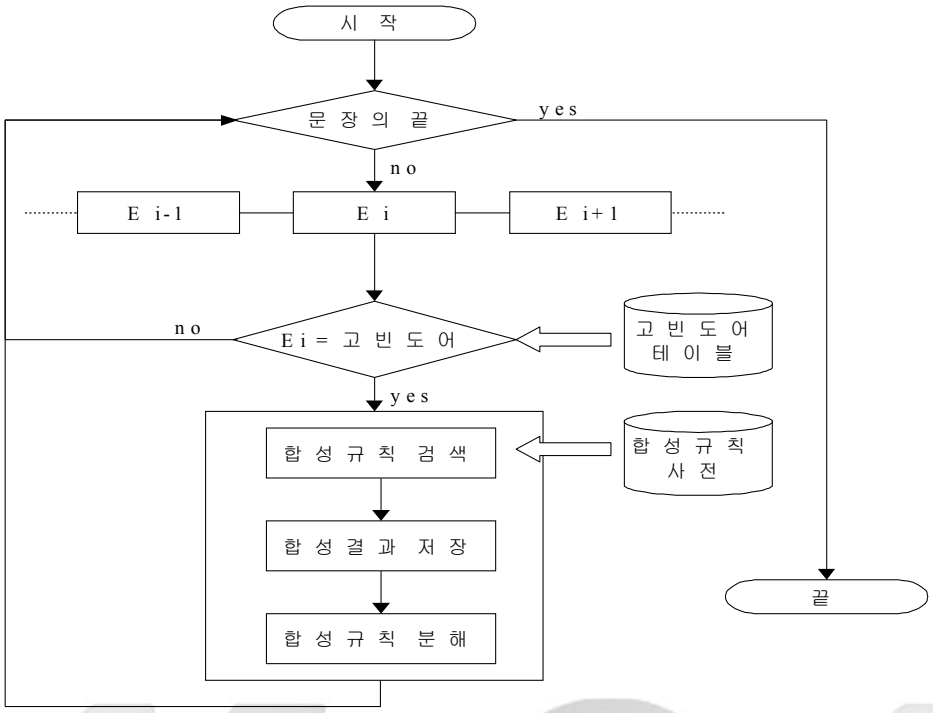
<그림 9> 고빈도어 추출 알고리즘

2. 합성명사 생성기

명사를 합성하는 이유는 광의의 단어가 색인됨으로써 검색시스템의 정확성을 떨어뜨리고 사용자의 탐색 시간을 늘린다. 본 연구의 합성명사 생성기는 기존 합성규칙[11]과 추가 제안한 규칙[12] 9가지에 따라 고빈도어를 대상으로 합성명사를 생성한다. <그림 10>은 합성명사 생성기의 흐름도이다.

---

11) 김민정, 한글 특성을 고려한 자동 색인기법(석사학위논문, 부산대학교 대학원 전자계산학과, 1993), p.35.
12) 김미진, 박미성, 최재혁, 이상조, 전게논문, pp.1125-1126.



<그림 10> 합성명사 생성기 흐름도

명사를 합성하는 방법은 형태소 해석 결과에서 한 단위(Ei)씩 읽어오면서, 합성 가능한 Ei가 발견되면 Ei를 중심으로 좌우를 살펴 합성규칙을 검사한다. 합성 규칙에 해당되면 명사를 합성하고, 분해한다. Ei의 좌우로 모두가 명사 합성에 성공했으면(if $comp_a$ and $comp_b$)양쪽을 붙여서 합성하고 분해한다. 이러한 과정으로 명사를 합성해 나가는 알고리즘은 <그림 11>과 같다.

```
[알고리즘 2] Noun_composition_method
Input    합성가능 단위          E1, ... , En
         고빈도어 추출결과      H1, ..., Hm
         합성규칙               C1, ..., Cp
Output   명사합성 결과          R1, ..., Rq

Noun_composition_method
for i = 0 to n do
   if (Ei.noun ∈ H) then
      if (is_복합명사(Ei.noun)
         save_noun_composition(E, C, i);

      // 합성규칙에 해당하면 명사를 합성, 분리
      for j = 1 to 2 do
         if (get_어절규칙(E, i-j, i) ∈ C) then
            save_noun_composition(E, C, i-j, i)
            compj = true;
         end if
         if (get_어절규칙(E, i, i+j) ∈ C) then
            save_noun_composition(E, C, i, i+j)
            compj+2 = true;
         end if
      enddo

      // 합성결과 좌우로 연결
      for a = 0 to 1 do
         for b = 2 to 3 do
            if (compa and compb)
               save_noun_composition(E,C,i-a, i+b)
         enddo
      enddo
   endif
enddo
```

<그림 11> 명사 합성 알고리즘

3. 합성명사 분해

추출된 색인어 후보들 중 고빈도어에 합성규칙을 적용하여 합성명사를 생성하기 때문에 합성규칙을 적용하지 않고 생성된 색인어를 찾아주지 못하는 경우가 있다. 그리고 합성규칙 적용으로 생성된 긴 합성명사의 경우에도 색인어로 사용되기 어려우므로 이러한 경우에 합성명사를 분해하여 단일명사 색인어를 만들게 되는데 분리의 기준을 두 가지로 세웠다. 하나는 조사를 기준으로 분해를 수행하고 또 하나는 고빈도어가 문서 내에서 합성명사의 일부분인 경우로 고빈도어를 중심으로 좌측인접명사 앞에서 분해를 하고, 또 우측인접 명사 뒤에서 분해를 해준다. 이상과 같이 고빈도어 색인어 추출기, 합성명사 생성기, 합성명사 분해의 세 과정이 완료되면 텍스트데이터베이스에 대한 색인데이터베이스의 구축이 완료된다.

이상의 전처리과정에서, 음성데이터베이스를 텍스트데이터베이스로 변환하였고, 변환한 텍스트데이터베이스로부터 색인데이터베이스를 추출하였는데, 이는 다양한 응용시스템에 활용될 수 있

겠지만 다음 장에서는 간단한 정보검색모델 제안을 통해 음성 질의나 문서 검색에 유용하게 사용될 수 있음을 보인다.

Ⅳ. 음성의 색인데이터베이스를 활용한 정보검색모델

본 장에서는 Ⅱ장의 음성처리 기반 기술과, Ⅲ장의 색인어추출 기반 기술로 완성한 색인데이터베이스를 정보검색의 전처리단계에서 활용 가능함을 보이기 위해 간단한 정보검색모델을 제안한다.

1. 정보검색

일반적으로 정보검색이란 자연언어로 구성된 문서의 내용과 명확하지 않은 사용자의 정보요구가 얼마나 근접한가를 계산하여 그 값이 높은 순서로 문서를 검색하는 것이다. 보통 문서의 내용과 사용자 질의의 유사도를 계산하는 정합(matching)과정의 효율을 위해서는 문서의 내용을 미리 분석하여 내부적인 표현으로 변환하는 것이 필요한데, 이를 색인(indexing)이라 한다. 색인과정에서 문서의 내용을 대표하는 용어를 선별하고 중요도를 계산하는데 비교적 간단한 형태소분석 수준에서의 자연어처리와 통계처리에 주로 의존한다.[13]

색인 추출은 정보검색에 가장 중요한 요소인데, 본 연구의 정보검색 모델에서는 이러한 색인추출을 위해 고빈도어를 이용한 색인데이터베이스를 이미 Ⅲ장에서 추출한 바 있고 이를 이용한다. 그리고 정보검색에 있어서 또 한가지 중요한 요소로는 사용자에게 제공할 질의 결과를 보여줄 때 적절한 문서와 적절하지 않은 문서를 선별해 사용자 요구에 부합되는 문서를 우선적으로 보여 주는 것이다.

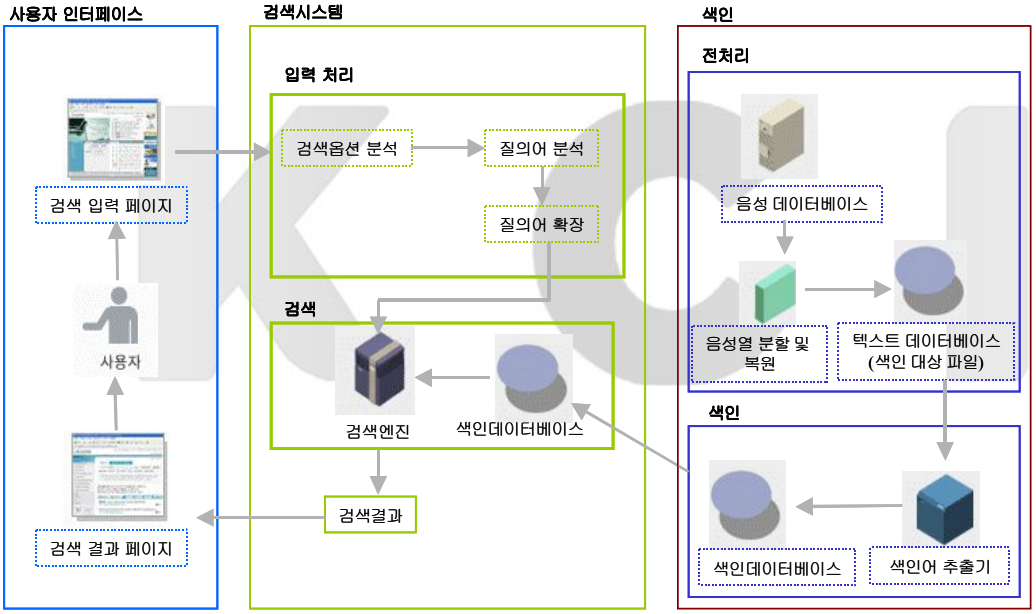
초기 정보검색모델에서는 문서의 우선순위를 보여주지 않는 불리언(Boolean)모델이 많이 사용되었다. 이 모델은 구현이 쉽고, 검색 시간이 짧으며, 간단한 불리언 연산자를 사용하여 비교적 쉽고 정확하게 질의를 표현할 수 있다는 장점이 있어서 사용자가 찾고자 하는 문서 집합을 이미 알고 있을 경우에는 불리언모델이 적합하기 때문에 상용데이터베이스나 서지검색시스템에서 널리 사용되어 왔다. 그러나 수기가 바이트 이상의 대용량 문서 데이터 중에서 적합한 문서를 검색하고자 할 때는 복잡한 불리언 질의를 사용해야 하기 때문에 불리언모델이 적합하지 않다. 이런

13) 맹성현, 주종철, "문서구조화와 정보검색," 한국정보과학회, 제16권, 제8호(1998, 8), pp.7-8.

경우 검색 결과에 대해 문서의 우선 순위를 매겨 상위 문서들을 사용자에게 제공하는 것이 가장 효율적이다. 보통 우선 순위에 따른 검색 결과는 사용자에게 필요한 정보를 얻는 시간을 최소화 시켜준다는 장점이 있다. 이처럼 문서에 우선 순위를 부여하는 것을 문서랭킹이라 하며, 문서랭킹을 계산하는 대표적 모델에는 퍼지 집합 모델, 확률 모델, 벡터 공간(Vector Space) 모델, 확장 불리언 모델, 지식 기반 모델 등이 있다.[14] 본 연구에서는 질의어와 문서에 대한 유사도를 비교하기 위해 전통적 방법인 백터 공간 모델을 채택한다.

2. 정보검색모델

본 연구에서 제안하는 음성데이터베이스의 색인데이터베이스를 활용한 정보 검색 모델은 <그림 12>와 같다. 시스템은 크게 사용자인터페이스, 검색엔진, 음성데이터베이스로부터 추출된 색인 부분으로 구성된다.



<그림 12> 음성데이터베이스의 색인데이터베이스를 활용한 정보 검색 모델

음성데이터베이스의 검색 과정은 다음과 같다. 전처리 단계에서 데이터베이스 관리자는 데이터

14) 김진숙, "정보검색시스템 KRISTAL-II에서의 우선순위 부여 모델 구현," *KORDIC Newsletter*, 제 15호(1999, 8), <http://www.kordic.re.kr/~news/letter/15/nl05.htm> [인용 2004. 8. 25]

베이스관리모듈을 사용하여 음성데이터베이스에 대한 텍스트 색인을 수행한다. 색인은 대량의 데이터를 빠르게 검색하게 해주므로 가장 결정적인 자료구조가 되는데, 본 모델에서는 내용기반 검색을 효율적으로 수행하기 위해 특정 단어를 가지고 있는 모든 문서를 빠르게 찾아내는데 가장 널리 사용되는 역파일(inverted file)구조를 사용한다.

<그림 12>의 오른쪽 모듈인 전처리 단계를 거쳐 색인구축이 완료되면 사용자 검색이 시작되는데, 사용자는 검색시스템에 자신의 정보 요구 즉 질의를 입력하면 시스템은 사용자 질의에 대해서 색인추출 방법과 동일한 기술을 적용하여 질의어를 분석하고 질의어 확장을 수행한다.

사용자 질의에 대해 검색된 문서들은 사용자에게 보내지기 전에 질의와 색인간의 연관성(likelihood)에 따라 순위화되며, 사용자는 유용한 정보 탐색을 위해 이 순위화된 문서를 검토한다. 여기서 사용자가 관심에 정확히 부합되는 문서를 선택하기도 하는데, 이것이 사용자 피드백(user feedback)이다. 이 과정을 수행하게 되면 시스템은 사용자가 선택한 문헌들을 대상으로 색인을 생성하고 질의를 변경하는데 이렇게 수정된 질의는 실제 사용자 요구에 더 가까운 표현이 된다.

1) 색인

정보검색을 위해서는 저장된 전문정보에 대하여 내용기반 색인이 우선 만들어져야 한다. 본 연구의 색인과정은 음성데이터베이스를 대상으로 전처리단계를 거쳐 변환된 텍스트데이터베이스가 색인 대상 파일이 된다. 색인 대상 파일은 효율적인 색인어 추출을 위해 출현 고빈도어를 대상으로 명사 합성 및 분해를 통해 색인작업을 수행한다. 여기서 생성한 색인은 음성데이터베이스에서 추출한 것임으로 검색옵션에 따라 검색 대상이 문서뿐만이 아니라 음성 즉 오디오 자료에도 적용 가능하다.

2) 사용자 인터페이스

정보 탐색자와 정보검색시스템이 대화하기 위해서는 사용자 인터페이스가 필요하다. 정보를 찾는 과정은 매우 불명확하여 사용자들이 정보시스템을 사용할 때 어떻게 정보에 접근할 것인가에 대해서 충분히 알지 못하기 때문에 정보 접근 시스템의 사용자 인터페이스는 원하는 정보를 표현하는데 도움을 줄 수 있다. 본 연구에서 제안한 검색 시스템의 정보 접근 단계는 다음과 같이 요약한다. 1) 사용자가 정보 요구를 가지고 질의어 입력한다. 2) 시스템은 질의어를 분석 및 확장한다. 3) 시스템으로 질의를 전송한다. 4) 검색엔진에서 처리된 결과를 받는다 5) 사용자는 시스템으로부터 받은 결과를 살펴보고, 평가하고, 해석한다. 더 적합한 정보를 얻기 위해 경우에 따라 피드백을 수행한다. 5) 탐색을 멈춘다. 6) 피드백할 경우에는 시스템이 질의를 변경해서 3) 단계를 재실행하게 된다.

여기서 중요한 부분은 사용자가 자신의 정보요구를 입력할 때 검색 옵션을 선택할 수 있게 하였다는 것이다. 검색 옵션에는 키워드와 음성에 대한 옵션을 제공한다. 키워드 옵션은 사용자가 입력한 질의에 대하여 색인 추출 기법을 이용하여 질의를 분석하고 확장하며, 음성에 대한 옵션은 키워드 입력 시 음성데이터베이스에 대하여 검색하는 옵션이다. 이러한 검색옵션을 선택하지 않았을 경우는 기본적으로 키워드로 검색을 수행하도록 하였다.

3) 검색시스템

사용자는 자신의 정보 요구를 키워드의 집합으로 제공하고, 검색엔진은 사용자의 질의에 가장 가까운 문서를 검색해 준다. 또 시스템은 선택된 연관성 측정방법에 의해 문서를 순위화한다. 순위화 작업은 사용자의 정보 요구를 만족시키는 결정적인 작업이다. 이러한 검색시스템이 수행하는 작업은 다음과 같다. 먼저 검색 입력 페이지에서 입력된 검색 옵션을 분석하고, 사용자가 입력한 질의어를 분석하는 과정과 질의어를 확장하는 과정을 수행 후 검색을 하게 된다. 검색결과는 사용자의 브라우저로 보내어 진다. 정보검색시스템에서 정보 자료를 검색하기 위해 사용되는 기법을 검색기법이라고 하고, 정보검색시스템에 주어지는 일정한 형식의 질의어를 검색문이라고 한다. 본 연구에서 정보검색기법으로 전통적인 벡터 모델 기법[15]을 사용한다. 불리언 모델이 질의나 문서의 키워드에 모두 이진 가중치를 할당하는데 비해 벡터 공간 모델은 질의나 문서의 키워드에 이진 값이 아닌 적절한 가중치를 할당할 수 있다. 그리고 키워드의 부분 매칭이 가능하며, 질의와 문서의 유사도에 따라 순위화를 할 수 있다. 벡터 공간 모델에서 모든 색인어는 서로 독립이라고 가정을 하며, 질의와 모든 문서는 식(7)과 같이 벡터 값으로 표현된다.

$$(7)\ d_i = (w_{i1}, w_{i2}, \cdots, w_{ik}, \cdots, w_{in})$$

식(7)에서 di는 i 번째 문서를, wik는 문서 di에서 나타난 k 번째 색인에 대한 가중치 값을 나타낸다. 문서 i에서 나타나지 않는 색인어에는 가중치가 0으로 할당된다. 문서 또는 질의 벡터들이 형성된 이후의 검색과정은 모두 벡터 연산에 의해 이루어진다. 본 연구에서는 문서 d와 질의 q사이의 유사도 측정을 위해 두 벡터 사이의 상관도를 구할 수 있는 코사인유사도(cosine similarity)공식[16]에 따라 값을 계산하여 문서와 질의의 유사성을 판단하도록 설계하였다.

---

15) Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval* (New York : ACM Press, 1999).

16) 상게서.

V. 결 론

지금까지 비정형 멀티미디어 데이터 유형인 음성데이터베이스를 텍스트데이터베이스로 변환하고, 변환된 텍스트데이터베이스로부터 색인데이터베이스를 구축함으로 텍스트와 음성을 대상으로 내용기반 정보검색 모델에 활용할 수 있음을 보였다. 음성데이터베이스를 텍스트데이터베이스로 변환하기 위해 우선 연속된 음성에 휴리스틱 정보와 어절 분할 정보 사전, 음절 분리 가능도와 같은 통계 정보를 이용하여 문서상의 어절에 근접한 단위로 분할하기 3단계의 분할 과정을 수행하였다. 그리고 분할된 단위를 대상으로 제안한 91개의 주 규칙과 240개의 부 규칙으로 구성된 음절복원규칙을 적용하여 텍스트데이터베이스로 변환한다. 올바른 텍스트데이터베이스 생성에 대한 검증을 위해 자연어 처리 기반 기술이 형태소분석기와 교정기가 사용되었다. 텍스트데이터베이스가 완성되면 이에 대한 효율적인 색인데이터베이스를 구축하기 위해 고빈도어를 이용한 색인어 추출기법을 제안하였다. 이 기법은 출현빈도가 상위 30-40%에 해당하는 명사들을 추출하여 이들을 대상으로 명사 합성 규칙과 분해 규칙을 통해 유용한 색인데이터베이스를 구축하였다. 그리고 이 색인데이터베이스는 음성을 기반으로 구축한 색인이므로 정보검색모델에서 검색옵션을 선택하게 하여 웹 상에 존재하는 문서 검색에서 뿐만이 아니라 음성 자료 검색에도 이용될 수 있음을 알 수 있었다. 향후 과제로는 제안한 정보검색모델의 구현을 통하여 외국 전문 멀티미디어 검색엔진개발을 선도하고 있는 싱잉피쉬, 스트림세이지 등에서 개발중인 음성검색시스템과의 비교를 통해 각 단계의 핵심 모듈에서 제안한 기능들을 조금씩 개선해 나가는 것이다. 그리고 제안한 정보검색모델의 사용자 인터페이스 부분 모듈을 수정하여 청각장애인들을 위한 시스템을 구현해 보거나 도서관에 보유하고 있는 오디오자료들에 대한 내용기반 검색 적용 및 이용자들의 음성 질의에 실제 적용시켜 보는 등 다양한 응용시스템들에 적용해 봄으로 시스템의 성능을 정확하게 검증해 보는 것이다.

<참고문헌은 각주로 대신함>