행정구역 위계정보와 편집거리를 이용한 오류입력에 강한 도로명주소 변환

송재용*
*고려대학교 컴퓨터정보통신대학원
e-mail:highfive@korea.ac.kr

Error tolerant Korean Roadname Address Conversion using Hierarchical Administrative Division and Edit Distance

Jae-Yong Song*
*Dept of Graduate School of Computer & Information Technology, Korea University

요 약

도로명주소가 법적 주소체계로 지정되고 2014년도부터 전면 시행을 앞두고 있는 상황에서 기존의 지번주소를 도로명주소로 변경하려는 수요가 늘고 있으며 그에 따라 주소 전환 서비스를 제공하는 솔루션들이 증가하고 있다. 행정구역 체계에 따라 단계별로 입력된 지번주소의 도로명주소로의 변환은 크게 어렵지 않고 변환 성공율도 상당히 높지만 자유롭게 입력하여 정제되지 않은 형태의 주소는 전환에 실패하는 경우가 많다. 본 논문에서는 전산입력된 지번주소를 도로명주소로 변환시 주소형태가 정제되지 않은 상황에서도 변환 성공률을 높이기 위해 행정구역 줄임, 일부 주소정보 누락, 오타 등 여러 가지 변형 케이스에 대해서도 유연하게 변환을 수행하는 방안을 연구하였다. 이를 통해 기존 지번주소의 표준 형태로의 정제는 최대 두 배까지 변환효율을 높일 수 있었다. 그러나 변환시 사용하는 도로명주소 매칭 테이블에 자료의 누락, 건물명의 불일치, 지번과 건물의 1:1 매칭이 되지 않는 경우가 존재하여 원활한 주소 전환을 위해서는 데이터의 정비가 필요하다.

1. 서론

2011년 7월 29일부로 일제 강점기부터 100년에 가까이 사용되어 오던 지번주소를 대신하여 도로를 기준으로 건물에 주소를 부여하는 도로명주소가 법적 주소체계로 지정되었다. 그러나 오랜 기간 사용되어온 기존 지번주소체계를 한 번에 바꾸기가 어려워 현재 기존의 지번주소와 병행 사용중이며 도로명주소만을 사용하는 전면 시행은 2014년으로 미뤄졌으나 이 기간이 얼마 안 남은 현 시점에서도 아직 활용도가 부족한 편이다. 또한 도로명주소 홈페이지 및 여러 관련업체에서 단건 및 다량의 주소전환 서비스를 제공하며 변환 솔루션 등을 판매하고 있으나 변환 대상이 되는 기존 지번주소가 올바르게 정제되지 않은 경우에는 변환 효율이 그리 높지 않다.

본 논문에서는 기존 지번주소의 체계 및 입력 오류 유형을 분석하고 변환효율을 높이기 위해 주소체계의 특징인 행정구역의 위계정보를 이용하여 검색효율 및 정확도를 높이고 행정구역 검색시 편집거리 계산을 적용하여 주소입력에 오타가 있는 경우에도 유연하게 대응을 수 있도록 하였다.

2. 기존연구

지번주소가 체계에서 주소 인식을 위한 연구는 과거에도 있어왔으며 주로 우편 시스템의 자동 분류를 위한 OCR 인식과의 연계를 위한 연구가 많았다. Kwanyong Lee et al[1]은 필기 입력주소의 행정구역을 단어별로 분리하여 은닉마르코프모델에 기반한 공기확률의 계산을 통해 인식율을 높이는 시도를 하였다. Gyeonghwan Kim et al[2]에서는 OCR을 통해 입력된 주소를 혼동행렬을 사용하여 인식율을 높이고 이를 행정구역체계 및 우편번호, 번지 및 건물정보를 이용해 주소의 순로코드를 발생시키는 방법을 제시하였다.

3. 지번주소의 구조와 입력유형 분석

3.1 주소변환의 예

기존에 사용되었던 지번주소의 기반이 되는 지번은 행정구역내 땅을 분할하여 그 번지를 부여한 체계로서 행정구역(법정동) 10자리, 산여부 1자리, 본번 4자리, 부번 4자리 등 총 19자리의 숫자로 구성된다. 여기에 아파트나 다세대 주택의 주소 구분을 위한 상세주소가 추가된다. 지번주소의 형식은 다음과 같다.

구분	위계	명칭	코드
행정구역 (법정동)	시도	서울특별시	11
	시군구	서초구	650
	읍면동	서초동	108
	리		00
산여부	산여부	(대지)	0
번지	본번	1436	0001
	부번	14	0001
상세주소		A 아파트 B동 C호	A 아파트 B동 C호

<표 1> 지번주소의 예

위와 같이 지번주소가 올바르게 정제된 경우 안전행정부 도로명주소 홈페이지(http://www.juso.go.kr)에서 제공하는 매칭 테이블을 통해 다음과 같이 도로명주소로 변환이 가능하다.[3]

구분	위계	명칭
행정구역 (법정동)	시도	서울특별시
	시군구	서초구
도로명		효령로68길
건물번호	본번	33
	부번	
상세주소		B동 C호(서초동, A 아파트)

<표 2> 도로명주소의 변환 예

이와 같이 지번주소가 정제된 상태이면 매칭테이블을 이용한 단순한 매핑으로 변환이 가능하다. 전산화된 시스템을 통해 입력된 주소정보는 입력시 UI를 통해 기존의 행정구역체계에 따라 입력할 수 있도록 강제되는 경우가 많아 입력시 주소정제가 되어 있으므로 변환 작업이 크게 어렵지 않으나 기존 수기로 작성된 주소정보를 그대로 전산파일로 옮기거나 행정구역 체계를 고려한 UI 가 없이 웹 페이지에서의 단건입력과 같이 사용자가 자유롭게 입력한 주소의 경우는 아래와 같이 여러 가지 오류가 존재할 수 있다.

3.2 지번주소 형태의 다양한 입력 예

도로명주소 홈페이지의 주소변환 요청 사례를 분석하여 다음과 같이 입력 형태별로 유형화 할 수 있다.

- 행정동 및 법정동의 혼용
- 약어 등 다양한 표기방식

법정동	변형표기
서울특별시	서울, 서울시
제주특별자치도	제주, 제주도
한강로1가	한강로제1동, 한강로1동, 한강로

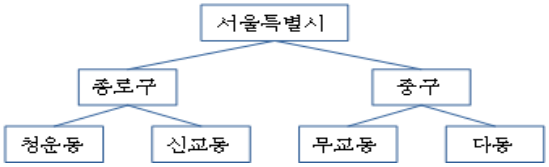
<표 3> 행정구역의 다양한 표기형태

- 구(舊) 행정구역 명칭을 그대로 사용
- 일부 행정구역 누락
- 띄어쓰기 및 오타

4. 지번주소의 정제

4.1 주소단위별 정제

행정구역정보를 Dictionary 및 Trie로 구성하고 위계정보를 조회할 수 있도록 다음과 같이 트리 형태로도 구성한다.



(그림 1) 행정구역 트리구성

또한 Dictionary에는 흔히 사용되는 변형/약어표기 및 구(舊) 행정구역 명칭을 alias로서 등록하여 변형입력에 대응할 수 있도록 한다. 단, 구 행정구역의 경우 단순 승격이나 명칭변경 등의 경우만 가능하며 행정구역의 분할 및 경계변경 등의 경우는 추적이 불가능하다.

행정구역은 어절단위로 분리하여 입력하는 것이 원칙이므로 정상적인 주소 입력의 경우는 Dictionary 검색만으로 정제가 가능하다. 만일 띄어쓰기 등의 오류로 검색이 불가능하면 음절 단위로 검색이 가능한 Trie를 이용하여 검색을 수행한다. Trie에서도 검색이 실패하면 행정구역 입력에 오타 또는 누락이 발생한 경우로서 다음 절에 기술할 편집거리 계산을 수행한다. 위의 각 검색단계에서 행정구역의 위계정보를 활용하여 전체 행정구역이 아닌 현재까지 검색된 행정구역에서 나올 수 있는 후보군을 대상으로 비교대상을 줄여 효율성을 높인다.

편집거리 계산 후에도 적절한 후보가 나오지 않을 경우 해당 위계의 행정구역은 누락되었다고 가정하고 다음 위계의 행정구역에 대하여 위와 같은 방식으로 진행한다. 이후 하위 행정구역을 찾았을 경우 위계정보를 이용하여 누락되었던 행정구역 정보를 복원시킨다.

지번은 (산) 0000-0000 형태로 구성되어야 하나 아래와 같이 산 번지의 누락, 본번/부번 구분의 기호의 다양한 표기 등의 형태가 존재하므로 이를 표준 형태로 정제한다.

표기형태	정제형태
132의 17	132-17
236번지(산번지의 경우)	산 236
27/3	27-3

<표 4> 지번주소의 여러 가지 표기형태

아파트나 다세대 등 하나의 토지 위에 여러 개의 주소를 부여하는 경우 정확한 건물을 구분하기 위한 단지/동/호 등의 상세주소가 추가적으로 필요하다.

또한 일반주택이 아닌 아파트나 빌딩 등의 건물은 관행적으로 지번을 누락하고 행정구역과 건물명만 표시하는 경우가 많아 상위 행정구역 및 건물명으로 누락된 지번 등을 복원해야 하는 경우도 존재한다.

건물명에서 오타가 발생한 경우도 편집거리 계산을 수행하여 가장 유사한 건물명을 찾는 것이 이론적으로는 가능하나 현재 도로명주소에서 제공하는 건물명이 건축물 대장에 등록된 건물명과 다른 경우가 많아서 안전행정부에서 건물명에 대한 정비작업 중인 바 본 연구에서는 건물명에 대한 편집거리 계산은 제외하였다.

4.2 편집거리 적용을 위한 행정구역 오류패턴 분석

주소입력에서 편집거리 적용을 위해서는 다음과 같은 행정구역의 특징 및 오류패턴을 수용해야 한다.

- 대부분의 행정구역은 길이가 3~4자로 비교적 짧다.
- 주로 한글로 구성되나 일부 숫자(~1동 등)가 포함된다.

- 다른 상위 행정구역에 속한 동일명칭의 행정구역이 흔하다. '중구'는 서울시 및 5개 광역시에 동일하게 존재한다.
- 행정구역의 명칭 중 숫자부분의 누락 또는 오기가 흔하다.
- 일부 행정구역의 글자가 누락되는 경우가 많지만 없는 글자가 추가되는 경우는 많지 않다.

이를 바탕으로 음소 및 음절간의 편집거리를 계산하기 위해 Kangho Roh et al[4]에서 제안한 GrpSim(그룹 기반의 유사도 알고리즘)에 기반한 알고리즘을 바탕으로 유사도를 측정하여 편집거리로 반영한다.

4.3 음소간 유사도 분류체계

음소간의 유사도를 적용하기 위한 분류는 Kangho Roh et al[4]에서 제안한 수정된 한글 음소 분류체계를 활용하여 다음과 같이 유사그룹을 분리하였다.

구분	초성	종성
그룹1	ㅁ,ㅂ,ㅃ,ㅍ	ㅁ,ㅂ,ㅄ,ㅍ
그룹2	ㄴ, ㄷ, ㄸ, ㅌ, ㄹ,ㅅ, ㅆ	ㄴ, ㄵ, ㄶ, ㄷ, ㄹ, ㄺ, ㄻ, ㄼ, ㄽ,ㄾ, ㅀ, ㅅ,ㅆ, ㅌ
그룹3	ㅈ, ㅉ, ㅊ	ㅈ, ㅊ
그룹4	ㄱ, ㄲ, ㅋ, ㅇ, ㅎ	ㄱ, ㄲ, ㄳ, ㅋ, ㅇ, ㅎ
그룹5	-	ㅅ

<표 5> 자음 유사도 그룹 분류체계

구분	모음종류
그룹1	ㅖ, ㅒ, ㅞ, ㅙ, ㅢ
그룹2	ㅟ, ㅚ, ㅔ, ㅐ
그룹3	ㅕ, ㅏ, ㅖ, ㅑ, ㅓ, ㅘ
그룹4	ㅡ, ㅜ, ㅗ, ㅠ, ㅛ

<표 6> 모음 유사도 그룹 분류체계

4.4 음소 및 음절의 유사도 정의

음소간 유사도 계산을 수행시 음소가 서로 일치할 경우는 양수인 M_p를 부여하고 음소가 삽입 혹은 삭제될 경우는 음수인 $-I_p$를, 서로 불일치할 경우 다른 유사도 그룹에 속할 경우 음수인 $-N1_p$, 같은 그룹일 경우는 음수인 $-N2_p$를 부여한다. 다른 그룹일 경우 차이가 더 커야 하므로 $-N1_p \leqq -N2_p$의 관계를 부여한다.

두 음소 a, b에 대해 각각의 음소가 속한 그룹을 g(a), g(b) 로 정의하고 두 음소간의 유사도 점수를 $\delta_p(a, b)$라고 정의하기로 하면 다음과 같이 정리할 수 있다.

- 음소 일치 : $\delta_p(a, b) = M_p$
- 음소 불일치 : $\delta_p(a, b) = -N1_p$, $g(a) \neq g(b)$
  $\delta_p(a, b) = -N2_p$, $g(a)=g(b)$,
  $-N1_p \leqq -N2_p$
- 음소 삽입/삭제 : $\delta_p(a, b) = -I_p$

음절의 경우를 정의하면 음소와 같이 일치, 불일치, 삽입, 삭제로 나누어 계산한다. 한글의 경우 음절이 일치할 경우 양수인 M_s을 부여하되 초,중,종성으로 이루어지는 한글의 특성상 개별 음소단위 합보다 큰 $M_s \geqq 3M_p$을 부여한다.

음절불일치는 음소단위로 다시 분리하여 모든 음소들이 다를 경우 음수인 $-N_s$를 부여하되 음절일치와 같이 음소단위 합보다 작은 $-N_s \leqq -3N1_p$을, 일부 음소가 일치할 경우는 각각의 음소간의 유사도를 합하여 계산한다.

음절 삽입은 음수인 $-I_s$ 를 부여하고 삭제의 경우는 $-D_s$를 적용한다. 음절의 추가는 전제조건에 따라 많이 발생하지 않으므로 불일치보다 작은 $-I_s \leqq -N_s$를, 삭제되는 경우는 더 잘 발생하므로 $-D_s \geqq -N_s$의 관계를 부여한다.

두 음절 A, B에 대해 두 음절간 유사도 점수를 δ_s라고 정의하기로 하면 다음과 같이 정리할 수 있다.

- 음절 일치 : $\delta_s(A, B) = M_s \geqq 3M_p$
- 음절 불일치 : $\delta_s(A, B) = -N_s, \leqq -3N1_p$, 모든 구성 음소가 불일치할 경우
  $\delta_s(A, B) = \sum_{k=1}^{3} (\delta_p(A^k B^k))$, A^1, A^2, A^3는 음절 A의 초성, 중성, 종성
- 음절 삽입 : $\delta_s(A, B) = -I_s \leqq -N_s$
- 음절 삭제 : $\delta_s(A, B) = -D_s \geqq -N_s$

숫자 등 한글외의 문자는 음소 분할을 할 수 없으므로 음절 단위의 삽입, 삭제, 변경만 수행 가능하다. '원효로2가'를 입력해야 하는데 '원효로'로 누락하는 경우나 '중계4동'을 '중계5동'등으로 숫자를 잘 못 입력하는 경우 등이며 한글에 비해 중요도가 낮다. 따라서 일치할 경우는 한글과 동일하게 M_s를 부여하고 삽입 또는 삭제의 경우 $-I_n$을 부여하고 음절 불일치는 그 두 배인 $-I_n$을 부여한다. 또한 $-I_n$은 한글 음절 삭제인 $-D_s$ 보다 큰 $-I_n \geqq -D_s$ 를 부여한다.

숫자로 구성된 음절을 C, 다른 음절을 D 라고 할 때 두 음절간의 유사도 점수를 δ_n라고 하면 다음과 같이 정리할 수 있다.

- 음절 일치 : $\delta_n(C, D) = M_s$
- 음절 불일치 : $\delta_n(C, D) = -2I_n$
- 음절 삽입/삭제 : $\delta_n(C, D) = -I_n \geqq -D_s$

4.5 행정구역 유사도 알고리즘

위 내용을 이용하여 두 문자열 A, B에 대한 유사도 SIM(A, B)의 계산을 위한 점화식을 다음과 같이 구성한다. A(i) 는 A 문자열에서 i번째 문자 음절을 뜻하며 H는 한글로 구성된 집합(H = {'가',...,'힣'})이다.

- Base Condition
  - $SIM(A(i), 0) = -i \times I_s$, for $|A| \geqq i \geqq 0$
  - $SIM(0, B(j)) = -j \times D_s$, for $|B| \geqq j \geqq 0$
- General Case
  - $SIM(A(i), B(j)) =$

$$\max\begin{cases} SIM(A(i-1), B(j)) + \begin{cases} -I_s, A(i) \in H \text{ and } B(j) \in H \\ -I_n, A(i) \notin H \text{ or } B(j) \notin H \end{cases} \\ SIM(A(i), B(j-1)) + \begin{cases} -D_s, A(i) \in H \text{ and } B(j) \in H \\ -I_n, A(i) \notin H \text{ or } B(j) \notin H \end{cases} \\ SIM(A(i-1), B(j-1)) + \begin{cases} \delta_s(A(i), B(j)), A(i) \in H \text{ and } B(j) \in H \\ -2I_n, A(i) \notin H \text{ or } B(j) \notin H \end{cases} \end{cases}$$

5. 실험결과 및 평가

행정구역의 유사도 알고리즘의 최적 파라미터를 구하기 위한 실험 및 실제 변환실험의 두 가지 실험을 수행하였다.

5.1 행정구역 유사도 최적 파라미터 설정실험

잘 못 입력한 행정구역명이 주어졌을 때 해당 행정구역과 동일 소속 위계에 있는 행정구역 명칭들과 편집거리를 계산했을 때 올바른 행정구역이 가장 높은 유사도를 가질 확률이 높을수록 좋은 방법이라고 판단할 수 있다. '남가자동'이라는 입력이 들어올 경우 서울시 서대문구의 16개 법정/행정동과 편집거리를 계산하여 '남가좌동'의 유사도가 가장 높을 경우 올바른 행정구역을 찾게 된다. 일치, 불일치, 삽입, 삭제를 나타내는 M_p, $N1_p$, $N2_p$, I_p, M_s, N_s, I_s, D_s, I_n등의 파라미터의 값을 주어진 제약조건을 만족시키는 범위 내에서 1~9까지의 수치를 변경하며 올바른 행정구역을 찾는 확률을 조사하였다. 전체 130개의 오류/정상 데이터셋을 기반으로 테스트한 결과 다음의 파라미터 조합이 가장 좋은 결과를 보였다.

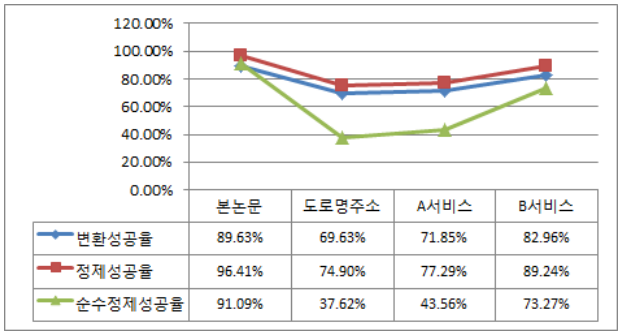
음소				음절					확률
M_p	$N1_p$	$N2_p$	I_p	M_s	N_s	I_s	D_s	I_n	92.31%
3	2	1	1	9	6	5	2	1	

<표 7> 최적 파라미터 값 설정

5.2 주소변환 실험

도로명주소 홈페이지에 변환요청 되었던 자료를 바탕으로 수집된 변환대상 지번주소자료 270건을 대상으로 안전행을 조사하였다.

상세주소등을 표시하는 법 등이 서로 약간씩 달라 행정구역 및 도로명, 건물 본번/부번까지 일치할 경우 올바르게 변환된 것으로 판정하였다. 그리고 지번주소와 도로명주소가 1:1 대응이 되지 않기 때문에(하나의 지번에 여러 건물이 존재하는 경우) 하나의 지번주소 변환 결과가 여러 개의 도로명주소로 나오는 경우도 변환결과의 목록이 정확하면 성공으로 판정하였다.



(그림 2) 주소전환 결과비교

변환 실패 결과 중 지번주소의 정제는 올바르게 이루어졌으나 도로명주소의 변환시 실패한 경우가 20건 가량 존재하였는데 이는 주소변환의 문제가 아니라 안전행정부에서 제공하는 매칭 데이터가 누락된 것이다. 이를 명확히 비교하기 위해 전체적인 변환성공율과 함께 지번주소의 정제가 성공한 정제성공율을 별도로 추출하여 비교를 수행하였고 이에 따라 전반적으로 성공률이 올라간 것을 확인할 수 있다. 마지막으로 이미 지번주소의 정제가 완료되어 단순 매핑을 한 자료를 제외하고 오류가 존재한 자료들의 정제 성공률을 조사한 결과가 순수정제성공율로서 성공률이 전반적으로 많이 하락하였으나 본 논문에 의한 결과는 기존 서비스에 비해 최소 약 20%, 최대 두 배 이상의 성공률을 보였다.

6. 결론

한글 기반의 편집거리 계산에 대한 연구는 이미 많이 이루어져 있으며 주로 검색 및 욕설 필터링과 같은 용도로 많이 사용되고 있다. 본 논문에서는 이러한 편집거리 알고리즘을 주소변환에 적용하고 행정구역의 위계정보 등을 활용하여 도로명주소 변환의 가장 핵심적인 지번주소 정제 성공률을 기존 대비 최대 2배 이상 높일 수 있었다. 일반 홈페이지의 회원 DB와 같이 행정주소체계에 맞춰 저장되어 있는 주소의 전환은 기존의 방법으로도 크게 문제가 되지 않지만 엑셀등을 이용하여 수동으로 주소를 관리하는 경우 또는 홈페이지에서의 주소검색 등 사람이 직접 입력하는 경우 등은 큰 효과를 볼 수 있다. 또한 주소검색뿐만 아니라 행정구역과 비슷한 형태의 위계정보를 가지는 자료의 매핑시에도 동일하게 적용이 가능하다.

다만 아직까지 주소변환을 위한 기반자료인 도로명주소 매칭 테이블의 자료가 누락된 경우가 있고 건물명 등이 건축물 대장 및 일반적으로 통용되는 명칭과 다른 경우가 존재하므로 도로명주소가 단일 법적 주소체계로 활용하기 위해서는 데이터의 빠른 정비가 필요하다.

참고문헌

[1] Kwanyong Lee, Jinwook Kwon, YiIlbyung Lee, "Korean Address Recognition using Syllabic Coocurrence Probability in Word Level", Dec, 1998

[2] Gyeonghwan Kim, Seokgoo Lee, Miyoung Shin, Yunseok Nam, "High-Speed Korean Address Searching System for Efficient Delivery Point Code Generation", Jun, 2001

[3] 안전행정부 주소정책과, "주소전환 및 정비 가이드", Jul. 2012

[4] Kangho Roh, Kunsoo, ParkHwan-Gue Cho, Sowon Chang, "Edit Distance Problem for the Korean Alphabet with Phoneme Classification System", Oct, 2010