우편번호 체계에서 사용중인 한글의 빈도수 조사

김 민기, 권 영빈
중앙대학교 컴퓨터공학과

A Frequency Measure of Hangul in Korean Zip Code

Min-Ki Kim and Young-Bin Kwon
Dept. of Computer Science & Engineering, Chung-Ang University

요 약

제약이 없이 자유롭게 쓴 오프라인 필기체 한글을 인식하는 문제는 응용분야에 따른 도메인의 정보를 이용함으로써 보다 쉽게 접근할 수 있다. 본 연구는 오프라인 필기체 한글 인식을 위한 한 도메인으로 우편봉투를 대상으로 하였을 때, 우편번호가 할당된 지명과 건물명을 대상으로 글자의 종류와 빈도수를 통계 분석하였다. 분석 결과 가능한 한글 조합 11,172자중 403자만이 쓰이고 있음을 알았다. 이러한 정보는 자소 분할이 어려운 오프라인 필기체 한글 인식에 있어, 문자 단위 정합을 사용했을때 인식속도 및 인식률 향상에 기여 할 것으로 생각된다.

I. 서 론

사회가 고도화 되어감에 따라 정보량이 폭발적으로 증가하게 되었고 이에 따른 정보의 수집 및 효율적인 관리를 위하여 자료를 자동으로 입력,처리,관리 및 출력하는 전산화가 필요하게 되었다. 그러나 컴퓨터를 비롯한 사무자동화 기기들의 발전에도 불구하고 자료 입력은 대부분 키보드를 통한 수작업으로 이루어지고 있는 실정이다.

문서의 자동 입력을 위해서는 문서 구조 분석과 문자 인식이 선행되어야 한다[13,14,15,16]. 국내의 문자인식, 특히 한글인식에 관한 연구를 살펴보면 문서의 자동 입력에 필요한 오프라인 인쇄체 인식에 대한 연구[4,5,6]와 최근 대두되고 있는 펜 컴퓨터에 활용되는 온라인 필기체 한글인식에 대한 연구[1,2]가 활발히 진행되어 상품화가

가능하거나 이미 상품화한 수준에 와있다. 그러나 오프라인 필기체 한글인식은 아직도 연구가 미흡한 상태이다.

오프라인 필기체 한글인식은 한글의 문자수가 방대하고 문자간의 유사성이 심할 뿐아니라 각 개인간의 필기 형태에 따른 변형때문에 상당한 어려움을 내포하고 있다. 따라서 오프라인 필기체 한글인식을 위해서는 다소 필기 형태에 제약을 가하거나 응용될 분야의 도메인 정보를 활용하는 방법[3,11,12]이 있다. 본 연구에서는 오프라인 필기체 한글인식의 응용중의 한 방법으로 우편봉투에 나타난 주소 인식을 생각하였다. 그리고, 우편번호 체계를 분석하여 나타나는 지명과 건물명을 구성하는 글자의 종류와 빈도수를 먼저 조사하게 되었다.

이 연구는 편지봉투의 인식에 따른 자동 분류를 위한 기초 연구의 일환으로 수행되었다. 우편번호 체계에 존재하는 글자의 종류 및 특성을 분석하여 인식기의 구성에 도움을 줄 수 있도록 하였다.

II. 우리나라 우편번호 체계

우편물의 구분작업을 능률적으로 처리하기 위하여 세계 각국은 우편번호제를 채택하여 행선지를 코드화하고 구분작업의 기계화를 추진하고 있다. 더우기 우편봉투내의 주소를 자동으로 인식할 수 있는 기기의 개발에도 많은 심혈을 기울이고 있어, 실용화된 제품도 많이 개발되어 판매되고 있다[9,10].

우편번호제는 1962년 서독에서 제도화된 것이 처음이며 우리나라에서는 70년 7월 1일을 기해 실시 되었다[7]. 처음 제정된 우편번호는 60년대말의 전국 주요 철도역을 기점으로 한 교통망에 기준을 두어 우편물을 배달하는 우체국에 3 또는 5자리 숫자를 부여한 것이었다. 그러나 70년 이후 경제의 고도성장에 따라 정보의 교환량이 급속도로 증가하여 이의 처리에 대응할 수 있는 우편번호 제도의 체계화가 요구되었다. 배달 우체국 단위로 부여된 그동안의 우편번호는 모든 주소를 나타낼 수 없기 때문에 우편물의 신속한 대량처리라는 측면에서 비효율적인 면이 많았다. 이러한 불편을 없애고 전산화에 적절히 대처할 수 있는 새로운 우편번호 제도가 제정되어 88년 2월 1일부터 시행되고 있다[7].

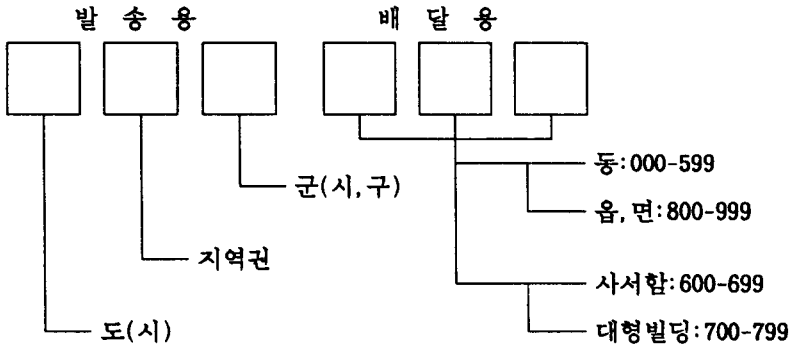
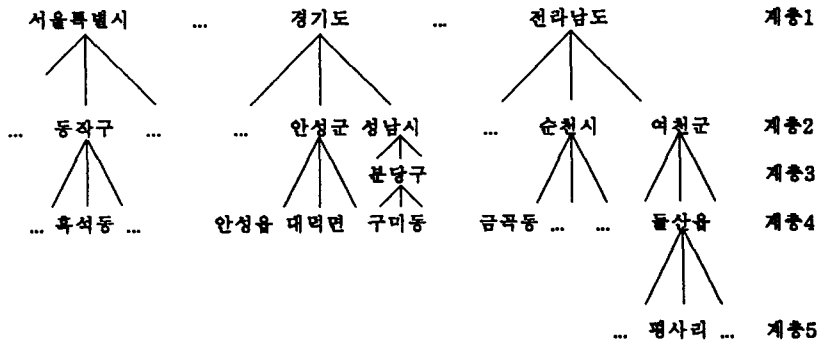


그림 1. 우편번호의 구성

우리나라의 우편번호는 발송용과 배달용으로 구분되어 도(시), 지역권, 군(시·구)의 번호가 3자리로 나열되어 있고 배달용으로 동·읍(면)을 3자리로 구분하여 전체가 6자리 숫자로 구성되어 있다(그림 1). 이 밖에 1일 평균 1,000통(면단위의 1일 평균 배달물량에 해당) 이상의 우편물이 배달되는 다량배달처(대형 빌딩)에는 고유번호가 부여되어 있다[8].

III. 우리나라 주소의 구성

우리나라 주소는 행정구역의 구조에 따라 계층적으로 이루어져 있다(그림 2). 최상위 계층은 도(특별시·직할시)이고 다음 계층은 군(시·구)이다. 구,시 아래에는 동이, 군 아래에는 읍(면)이 존재하나 성남시 분당구, 안양시 만안구 등 규모가 커진 시는 그 아래 구를 두고 있다. 다음으로 동·읍(면),리(동)으로 이루어져 있다.



(a) 일반적인 경우 (b) 계층3을 포함하는 경우 (c) 계층5를 포함하는 경우

그림 2. 계층적 주소 체계

우편물 자동 분류를 목적으로 주소를 분류하면 크게 둘로 나눌수 있다. 주소 전반부는 계층1과 계층2로 이루어지며 성남시, 안양시 등 7개(92년 10월 1일 현재)의 규모가 큰 시는 계층3을 포함하게 된다(그림 2-(b)). 주소 후반부는 계층4로 이루어져 있으나 배달상의 편의를 위해 별도의 우편번호가 할당되는 경우는 계층5를 포함하게 된다(그림 2-(c)). 주소 전반부는 우편번호 상위 3자리 발송용에 대응되고 주소 후반부는 하위 3자리 배달용에 대응된다.

IV. 글자의 종류와 빈도수 조사

주소에 나타나는 글자의 종류와 빈도수를 조사하기 위해서 먼저 체신부에서 발행한 우편번호부[7]를 기초로 우편번호 사전을 구성하였다. 사전은 주소 체계에 따라 계층적으로 설계하였고 사서함에 관한 우편번호는 별도로 구성하였다.

주소 전반부는 계층1,2,3을 하나로 했으며, 주소 후반부는 우편번호 하위 3자리가 할당되는 동·읍(면), 사서함, 대형 빌딩을 기준으로 하고 우편 배달의 편의성에 따라 낙도, 오지와 같은 지리적 여건으로 별도의 우편번호가 할당되는 예외적인 경우는 리(동)을 포함시켰다.

오프라인 필기체 한글 주소 인식에 활용할 목적으로, 주소 전반부와 후반부에 나타나는 글자를 대상으로 종류와 빈도수를 조사하였다. 주소 전반부에 나타나는 글자를 대상으로 조사하여 <표 1>의 결과를 얻었다. 주소 후반부는 지명, 건물명, 사서함으로 나누어 분석하였다<표 2 : 사용되는 글자의 총괄표는 부록1. 참조>. <표 1>과 <표 2>에는 내용별로 나누었을때 중복하여 나타나는 것을 모두 표현하였다. 물론 <표 1>의 각 항에 나타나는 글자들은 <표 2> '전체'란의 각 항에 나타나는 글자들에 모두 포함된다. 그 결과 가능한 한글 조합 11,172자 중에서 우편번호가 할당된 지명과 건물명에 사용되는 글자는 총 403자에 불과하다는 것을 발견하였으며 음절별로 나타날 수 있는 글자의 수효도 각기 다른 것을 알게 되었다. 또한 우리나라 우편번호 체계에서 나타난 403자를 한글의 6 타입으로 분류 하여 빈도수를 표시하면 <표 3>과 같이 나타나고 있다.

표 1. 주소 전반부에 나타나는 글자

첫 음절에 나타나는 글자	106자
둘째 음절에 나타나는 글자	79자
나타나는 모든 글자의 수효(중복제외)	143자

표 2. 주소 후반부에 나타나는 글자

	지 명	건물명	사서함	전 체
첫 음절에 나타나는 글자	290자	117자	55자	304자
둘째 음절에 나타나는 글자	276자	119자	47자	294자
나타나는 모든 글자의 수효 (중복제외)	355자	246자	106자	403자

표 3. 403자에 대한 Type별 분류

한글 6 Type						
빈도수	58개	169개	42개	102개	19개	13개

V. 결 론

주소에 나타나는 글자의 종류와 빈도수를 조사해본 결과 우편물을 발송지별로 분류하기 위해서는 주소의 전반부 만을 인식하면 되기 때문에 143자의 한글을 인식하면 되고, 배달지별로 분류하기 위해서는 403자만을 인식하면 됨을 알았다. 신도시나 새로운 건물등이 생겨난다 하더라도 예외적인 경우를 제외하고는 기존의 지명, 건물명에 나타나는 글자들을 사용할 것으로 예측되기 때문이다.

자소 분할이 어려운 오프라인 필기체 한글 인식에 있어 문자 단위 정합은 자소 분할이 필요없고 필기 형태에 대한 제약을 가하지 않기 때문에 가장 일반적인 방법으로 볼 수 있다. 그러나, 정합해야 할 부류의 수가 많아 자소 단위 정합이나 영역 단위 정합에 비해 상대적으로 인식 속도가 느리고 높은 인식률을 기대할 수 없었다.

오프라인 필기체 한글 인식의 한 도메인으로 우편봉투를 대상으로 할때 상기에 조사된 결과를 활용하여 문자 단위 정합을 수행한다면 인식속도는 물론 인식률을 향상시킬수 있을 것으로 생각된다. 또한 계층적으로 설계된 사전 정보를 활용하고 우편번호를 이용하여 후처리를 수행할 경우 오인식 및 미인식을 크게 줄일 수 있을 것이다.

VI. 참고문헌

[1] 권 오성, 권 영빈, "동적인 선분생성을 이용한 온라인 한글 필기 인식", 한국정보과학회 봄 학술발표 논문집, 제 20권 1호, pp. 151-154, 1993년 4월.

[2] 신 봉기, 김 진형, "통계적 방법에 의한 온라인 한글 필기 인식", 제4회 한글 및 한국어 정보처리 학술발표 논문집, pp. 533-542, 1992년 10월.

[3] 이 성환, 박 정선, "통계적 특징 추출 방법을 이용한 샘물체 필기 한글의 오프라인 인식", 제4회 한글 및 한국어 정보처리 학술대회 논문집, pp. 237-248, 1992년 10월.

[4] 최 동혁, 류 성원, 강 현철, 박 규태, "계층 구조 신경망을 이용한 한글 인식", 대한전자공학회 논문지, 제 28권 B편 제 11호, pp. 1-7, 1991년 11월.

[5] 김 화룡, 방 승양, "네오코그니트론 방식의 신경 회로망을 이용한 한글 문자 인식", 한국정보과학회 논문지, 제 18권 제 3호, pp. 280-289, 1991년 5월.

[6] 정 주성, 김 춘석, 박 충규, 황 희용, "윤곽선 추적에 의한 고딕체 한글의 신속 인식에 관한 연구", 전기학회 논문지, 제 37권 제 8호, pp. 579-587, 1988년 8월.

[7] 학원세계대백과사전(전32권), 제 22권, pp. 437-441, 학원출판사, 1993년 1월.

[8] 체신부, "우편번호부(1992년 10월 1일 현재)", pp. 2-63, 1992.

[9] Mauricette Feuillas, Serge Hugot, Emmanuel Miette, "Architecture of a Multigrey Level Videocoding System", JET POSTE 93, pp. 495-502, June 1993.

[10] T. Ono, T. Shima, M. Yoshioka, "Automated Material Handling Systems in 'Osaka Letter Post Office' and 'Osaka Parcel Post Office'", JET POSTE 93, pp. 465-472, June 1993.

[11] Michel Gilloux, Jean-Michel Bertille, Manuel Leroux, "Recognition of Handwritten Words in a Limited Dynamic Vocabulary", JET POSTE 93, pp. 148-155, June

1993.

[12] Venu Govindaraju, Ajay Shekhawat, Sargur N. Srihari, "Interpretation of Handwritten Addresses in US Mail Stream", IWFHR III, pp. 197-206, May 1993.

[13] Kyoon Ha Lee, Kie-Bum Eom, R. L. Kashyap, "Character Recognition Based on Atttribute-Dependent Programmed Grammar", IEEE Trans. on PAMI, Vol 14, No. 11, pp. 1122-1128, Nov. 1992.

[14] Sargur N. Srihari, "High-Performance Reading Machines", proceedings of the IEEE, Vol. 80, No. 7, pp. 1120-1132, July 1992.

[15] Shuichi Tsujimoto, Haruo Asada, "Major Components of a Complete Text Reading System", proceedings of the IEEE, Vol. 80, No 7, pp. 1133-1149, July 1992.

[16] Joachim Kreich, Achim Luhn, Gerd Maderlechner, "An Experimental Environment for Model Based Document Analysis", ICDAR 91, pp. 50-58, 1991.

부록 1. 우편번호 체계에서 사용중인 한글의 빈도수

동	4596	면	1389	대	520	산	497	가	449	리	394	서	357	신	355	사	285	남	281
성	281	학	281	천	280	도	237	곡	229	정	229	읍	209	원	205	화	202	문	201
국	200	교	199	전	193	수	186	평	181	양	179	장	178	부	169	송	169	암	163
우	163	안	157	상	156	용	147	북	146	중	145	지	143	금	141	내	140	봉	136
구	134	덕	134	월	134	주	134	현	131	포	130	하	125	청	121	삼	121	로	118
함	117	체	116	건	108	방	106	촌	106	계	105	림	103	이	102	호	102	딩	100
빌	100	물	99	광	97	명	94	창	94	고	92	영	90	석	89	연	87	일	87
의	86	진	84	홍	84	인	82	공	81	한	81	오	80	당	77	기	73	운	73
소	70	선	69	자	69	미	66	보	66	제	66	경	64	관	63	강	62	마	58
두	57	유	56	목	52	앙	52	옥	51	무	50	울	50	태	50	풍	50	죽	49
아	48	외	48	회	48	여	47	초	45	백	44	충	44	량	42	해	42	노	41
황	41	시	40	항	40	개	39	업	38	은	38	조	38	룡	37	만	37	본	37
비	37	군	36	내	36	행	36	효	36	반	35	북	35	종	33	모	32	단	31
등	31	감	30	좌	30	법	29	야	29	흥	29	과	28	세	28	직	28	왕	26
길	25	점	25	거	24	농	24	역	24	임	24	망	23	범	23	음	23	치	23
간	22	병	22	탄	22	갈	21	래	21	생	21	실	21	쌍	21	어	21	파	21
합	21	다	20	달	20	라	20	예	20	적	20	통	19	낙	18	릉	18	민	18
온	18	육	18	둔	17	락	17	류	17	향	17	심	16	율	16	철	16	칠	16
토	16	후	16	괴	15	불	15	센	15	스	15	위	15	잠	15	타	15	검	14
궁	14	귀	14	능	14	배	14	순	14	십	14	재	14	저	14	캠	14	퍼	14
담	13	둑	13	륜	13	웅	13	팔	13	근	12	례	12	설	12	승	12	율	12
추	12	춘	12	령	11	척	11	혹	11	희	11	각	10	답	10	발	10	와	10
완	10	나	9	막	9	변	9	속	9	응	9	입	9	작	9	축	9	탑	9
승	8	악	8	약	8	익	8	총	8	판	8	격	7	김	7	녹	7	랑	7
려	7	손	7	술	7	압	7	염	7	증	7	차	7	환	7	논	6	련	6
묵	6	번	6	별	6	찰	6	채	6	데	5	록	5	료	5	묘	5	박	5
벌	5	벽	5	침	5	험	5	훈	5	갑	4	권	4	누	4	돈	4	롯	4
린	4	빈	4	색	4	식	4	요	4	잔	4	카	4	택	4	톨	4	퇴	4
패	4	필	4	할	4	협	4	횡	4	곤	3	곶	3	극	3	납	3	낭	3
녕	3	돌	3	드	3	득	3	력	3	릭	3	립	3	밀	3	분	3	빙	3
섬	3	언	3	엄	3	절	3	출	3	코	3	키	3	터	3	형	3	활	3
휘	3	휴	3	견	2	결	2	골	2	콧	2	규	2	균	2	그	2	난	2
냉	2	눌	2	니	2	력	2	렴	2	론	2	롱	2	말	2	삽	2	새	2
쇼	2	숙	2	애	2	족	2	줄	2	표	2	피	2	핑	2	헌	2	혜	2
걸	1	겨	1	겸	1	괘	1	굴	1	궐	1	날	1	녈	1	넓	1	뉴	1
느	1	늑	1	둑	1	등	1	란	1	랜	1	럴	1	레	1	륙	1	맹	1
몽	1	밤	1	붕	1	삭	1	살	1	쇄	1	쇠	1	알	1	앵	1	얄	1
열	1	엽	1	울	1	욤	1	웅	1	왜	1	육	1	욱	1	워	1	윈	1
윤	1	접	1	존	1	집	1	징	1	책	1	처	1	취	1	측	1	콤	1
탁	1	탕	1	텔	1	트	1	특	1	팽	1	품	1	프	1	픽	1	헤	1
혈	1	홀	1	홀	1														