시뮬레이션 기반의 한글 성명 유사도 산출 알고리즘의 최적 가중치 산정 방법

정병희*, 이규호**, 박동하**, 최영환**, 양준용*1)
*비알씨(주)
**GE헬스케어코리아
e-mail:buxbany@gmail.com

Estimate weighted value for korean name similarity computing algorithm based on simultation.

Jeong, Byung-Hui*, Kyoo-Ho Lee**, Dong-Ha Park**, Young-Hwan Choi**,
Yang, JunYong*
*Bio Research Complex
**Ge Healthcare Korea Technology Center

요 약

국내 MPI 시스템의 도입을 위하여 한글성명에 대한 유사도 비교 알고리즘이 필요하다. 기존의 영문 성명 비교 알고리즘의 경우 조합형 글자를 지원하지 않기 때문에 한글에 적용할 경우 좋은 결과를 내지 못한다. 이러한 문제를 해결하기 위해 한글성명 매칭 알고리즘을 연구하였으며 본 논문에서는 한글 유사도 알고리즘에서 사용되는 여러 가중치의 최적 값을 시뮬레이션을 통해 산정하는 방법에 관하여 연구하였다.

1. 서론

현재 병원 별로 고유의 병원정보시스템이 운영되고 있어 병원 내 의료서비스 효율이 많이 향상이 되었다. 그러나 병원마다 운영되는 병원정보시스템의 수집 정보 및 포맷이 다르기 때문에 병원 간 의료정보 교류가 어려워 환자는 같은 질병으로 다른 병원을 찾아 진료를 받을 경우 처음부터 검사 및 진단을 다시 받게 되어 여러 가지 시간적 금전적 낭비가 발생되고 있다.

미국과 같은 선진국에서는 MPI(Master Patient Index) 알고리즘이 연구 개발되어 서로 다른 병원 간 환자정보를 비교 분석하여 사용할 때 환자 개인을 식별할 수 있도록 사용되고 있고, 이를 통해 병원진료 서비스의 질을 높이는 동시에 시간적, 금전적 낭비를 줄이고 있다.

MPI 시스템의 국내도입을 위하여 한국에 맞게 변경이 필요한 부분들이 있다. 예를 들어 한글성명의 경우 알파벳을 이용한 서양방식의 비교 알고리즘을 사용할 경우 높은 낮은 매칭률을 보여준다.

이러한 문제를 해결하기 위하여 한국인의 성명 특징을 이용한 매칭 알고리즘에 관한 연구를 진행하였다 하지만 기존 연구에서는 가중치의 값을 개발자의 통찰을 기반으로 적정한 값을 대입하여 사용하였다. 본 논문에서는 시뮬레이션을 통해 기존 가중치를 넘어 최적의 가중치를 찾는 방법에 대해 연구를 진행하였다.

2. 한글성명 매칭 알고리즘

기존 한글유사도 평가 알고리즘들의 경우 외래어와 비속어에 관한 유사도 연구가 주를 이루었다[1]. MPI시스템에서는 한글성명의 유사도를 비교하는 알고리즘이 필요하여 한글성명에 관한 특징을 추출하여 반영하였다.

특화된 알고리즘은 통계청 데이터를 기반으로 한국인의 한국성씨에 속하는 경우 가중치 값을 부여하는 것과 한글의 두음법칙, 발음적 특성과 같은 한글특성을 반영하고 키보드의 배치를 기반으로 자주 발생할 수 있는 오타 정보를 이용하여 유사도 그룹을 생성하고 같은 유사도 그룹에 속할 경우 가중치 값을 부여하는 방법, 많은 양의 성명을 분석하여 자주 사용되지 않는 음소를 추출하여 해당 음소와 유사한 그룹을 제공하여 해당 그룹에 속할 경우 가중치 값을 부여하는 방식을 이용하였다.

이러한 한글성명에 특화된 알고리즘 설계를 통해

1) 교신 저자 : junyong0125@gmail.com

* "본 연구는 보건복지부 보건의료연구개발사업의 지원에 의하여 이루어진 것임(A112020)."

추출된 가중치는 표1과 같고 총 9개로 이루어져있다. 이중 일치했을 때는 긍정적인 값을 불일치했을 경우에는 부정적인 값을 배정하여 변별력을 높일 수 있다.

<표 1> 한글 유사도에서 사용되는 가중치값 들

표현	설명(2개 입력스트링 비교)
M	음소 a, b가 일치할 경우
Na	음소 a, b 가 유사음소쌍 테이블 내 같은 그룹에 속하지 않은 경우
Nb1	음소 a, b 가 유사음소쌍 테이블 내 속하는 경우 (유사음소 쌍 별 차등 가중치 적용)
Nb2	
Nb3	
La	음소 a, b가 저 빈도 대체 음소 테이블 내에 속하지 않는 경우
Lb	음소 a, b가 저 빈도 대체 음소 테이블에 속하는 경우
I	음소가 삽입되거나 삭제된 경우
F	(성씨 매칭) 성씨음절 A, B 둘 중 하나라도 성씨 통계에 포함되지 않은 경우

기존 한글 성명 유사도 알고리즘에서는 개발자가 가중치의 사용목적과 가중치의 중요도에 따라 적절하다고 생각한 값을 적용하여 사용하고 있다.

3. 가중치 최적화를 위한 시뮬레이션 방법

기존의 한글 유사도 논문들에서 사용하던 최적 가중치 산출 방식[2]을 기반으로 최적의 가중치를 산출 할 수 있는 시뮬레이션 방법을 한글성명 유사도 산출 알고리즘에 적용하였다. 한글성명 유사도에서 사용되는 가중치는 총 9개로 이루어지며 그중 긍정적인 의미의 가중치인 M, Nb1, Nb2, Nb3, Lb의 경우에는 0부터 10까지의 값을 부정적인 의미의 값인 Na, La, I, F의 경우 -5부터 5까지의 순차적으로 변경해가며 시뮬레이션을 진행하도록 한다. 시뮬레이션 결과의 좋고 나쁨을 판단하기 위하여 한글성명 데이터세트를 준비하고 먼저 유사성명 데이터세트를 이용하여 유사 성명쌍을 96% 이상으로 유사하다고 판단하는 유사도를 기준으로 비유사 성명데이터세트와의 유사도 비교 결과 값이 유사성명을 통해 구한 유사도를 넘는 경우 이를 에러율로 보고 에러율을 통하여 유사도의 최적정도를 판단하기로 한다.

4. 실험

가중치 최적화를 위한 시뮬레이션 방법에서 제안한 시뮬레이션을 진행하기 위하여 매뉴얼하게 작업한 유사 성명 1000개쌍과 비유사 성명 30만개 쌍을 이용하여 시뮬레이션을 진행하였다.

실제 적용하는 병원 시스템에서의 성능을 판단하기 위하여 협력관계의 병원 시스템에서 성명 데이터를 제공받아 성명 데이터세트를 구현하였다.

위 데이터 세트의 시뮬레이션 결과 최적의 결과 3개의 조합을 추출한 결과는 표2에서 보여주고 있다.

<표 2> 시뮬레이션 결과

표현	조합1	조합2	조합3
M	6	7	7
Na	-1	-3	-1
Nb1	4	4	6
Nb2	2	2	4
Nb3	1	1	2
La	0	-2	-1
Lb	2	4	3
I	-2	-3	-5
F	-4	-3	-6
에러율	0.54%	0.67%	0.71%

5. 결론

실험을 통해 시뮬레이션의 결과 최적의 결과값은 조합1의 값인 M, Na, Nb1, Nb2, Nb3, La, Lb, I, F가 6, -1, 4, 2, 1, 0, 2, -2, -4일 때 0.54%로 가장 적은 에러율을 보여준다.

이를 통하여 기존 시스템에서 개발자의 통찰을 기반으로 적용된 가중치를 이용할 경우의 에러율이 0.94% 인 것을 가만할 때 최적값을 적용했을 때 에러율이 0.54%인 조합을 찾아내어 시뮬레이션을 기반으로 최적의 가중치를 산출 해냈다고 판단 할 수 있다.

이를 기반으로 MPI 시스템에서의 한글성명 지원에 개선을 예상 할 수 있으며 지속적인 시스템 고찰을 통하여 지속적으로 개선해 나갈 예정이다.

참고문헌

[1] 고숙현, 문맥을 고려한 유사 외래어 검출 알고리즘, 충북대학교

[2] 노강호, 박근수, 조환규, 장소원, 음소의 1차원 배열을 이용한 한글 유사도 및 편집거리 알고리즘, 정보과학회논문지 : 컴퓨팅의 실제 및 레터 제17권 제10호, pp. 519-526, 2011.