데이터베이스상의 한글 자모단위 비교를 통한 데이터 정정기법

김대환[O] 백두권
고려대학교 컴퓨터과학기술대학원, 고려대학교 컴퓨터학과
dhwankim@korea.ac.kr, baik@swsys2.korea.ac.kr

A Revising Method using Phoneme Comparison for Databases with Korean Character Set

Daehwan Kim[O], Doo-Kwon Baik
Computer Science Technology Graduate School, Korea University
Dept. of Computer Science & Engineering, Korea University

요 약

코드로써 관리되어있지 않은 데이터베이스 내의 다양한 속성들이 시간이 흐름에 따라 정보로써 가치를 갖게 되면서, 비코드성 한글 데이터의 정형화에 대한 요구가 증가하고 있다. 정형화에 있어 한글의 특수성 중에 하나는 한글자료의 경우 KSC5601, CP949등을 사용하여 음절단위의 문자셋을 사용하여 음절단위로 저장 관리한다. 그런데 입력 시점에서는 자판기등을 이용하여 음소단위로 데이터를 입력하면서 발생하는 오류 및 비정형 데이터의 유입의 문제 등을 내포하고 있다. 이러한 문제를 해결하기 위하여 데이터의 저장단위인 음절이 아닌 음소 단위의 비교를 통하여 데이터를 정정하는 기법을 제안하고자 한다.

1. 서 론

더밍박사가 품질관리라는 개념을 전산분야에 적용하면서 데이터의 품질에 관한 문제가 인식되기 시작하였다.[1,7,8] 그 이후 데이터 품질 관리의 일부로써 데이터의 정정이라는 부분에서 데이터의 정형화와 데이터의 강화라는 개념을 적용하여 데이터를 필요한 형태로 유지 관리하기 위해 지속적인 연구가 되어 왔다. 그러나 이러한 연구는 주로 문자의 경우 음소단위를 최소 문자셋으로 사용하는 Roman Character Set을 기준으로 연구되어 오고 있어 음절단위의 문자인 완성형 한글데이터의 경우에는 적용에 어려움을 겪고 있는 현실이다. 따라서 한글 데이터의 정정을 위해서는 영문 데이터의 정정을 위한 연구를 고찰하여 한글셋에 맞도록 조정을 하여야 한다. 일반적으로 정정에 있어 도메인 값의 정형화 및 숨겨져 있는 Hidden Value 또는 의미를 찾는 강화작업에 있어 어떠한 방법들이 제안되어 있는지를 파악하고 이를 완성형 한글셋에 적용 및 기법의 효율성을 향상 시킬 수 있는 방안을 제안하고자 한다.

본 논문의 구성을 보면, 2장에서는 데이터베이스상의 데이터 정정에 적용된 기법 및 영문데이터를 처리를 위한 상용제품들의 접근방법을 기술하였으며, 3장에서는 한글 데이터의 정정을 위한 한글 데이터 내에 존재하는 오류가 어떤 것이 있는지 오류를 분류하고, 오류를 정정할 수 있는 방안에 대해 설명하였으며, 4장에서는 한글데이터의 오류 정정을 위한 유사도 검사를 위한 유사성 함수, 오류 정정의 주요 항목 및 오류수준의 측정, 데이터의 정형화를 위한 한글 데이터의 정정기법에 관하여 논하고 있다.

2. 데이터베이스상의 데이터 정정에 관한 연구

데이터베이스상의 데이터의 Well-Formed Data로 관리하기 위하여 RDBMS의 경우 E-R 모델링 단계에서부터 정규화를 통하여 데이터의 중복성을 제거하고 정형화된 데이터로 관리하기 위해 노력을 하고 있으나, E-R 모델링이 산업계에서 적용되기 시작한 1980년대 이후 지속적으로 관리 기법들이 개발되어 왔다.[2,5]

코드화가 어려운 자유형 데이터에 대한 데이터의 정형화를 위하여 David Loshin은 도메인 값의 정정시 유사음 수준 측정을 통한 결정론적 방법론을 적용하여 데이터의 정정기법을 적용하였다. 이러한 정형화는 영문데이터 처리에 목적을 두고 있다.

데이터 정정을 위한 상용제품으로는 Trillium Software사의 Trillium System과 Ascential Software사의 Integrity와 같은 영문 데이터 품질관리 Tool의 경우 성명과 주소 정정에 목적을 두고 있다.

고객명의 분류 :고객명 필드를 Semi Structure를 이용하여 성명의 직함, 성, 이름 등의 구분

주소 분류 :각각의 주소 항목의 수준에 맞게 분류

주소정보 표준화:미국 주소체계에 맞도록 자유형 데이터를 가공하여 주소 데이터를 정형화 데이터의 오류를 정정하기 위해서는 데이터의 오류를 찾아내어 데이터 상태를 측정하고 다양한 표준화 규칙에 맞추어 정형화한 이후에, 데이터 내에 의미적으로 숨겨져 있는 추론 가능한 정보를 찾아내어 데이터를 재구성하는 강화에 주안점을 두고 있다.[2,8] 이와 같은 일련의 작업을 통하여 오류의 제거 및 데이터의 중복성을 제거하게 보다 유용한 데이터의 확보를 목적으로 하고 있다.

3. 한글데이터의 오류 유형

인류가 만들어 사용해 온 글자는 단어문자, 음절문자, 음소문자의 3종류가 있으며, 단어문자는 단어의 수효만큼, 음절문자는 음절의 수요만큼 글자가 있어야 하고, 음소문자는 음소의 수요만큼 글자가 있어야 한다. 음소 수보다 글자 수가 많으면 읽기는 문제가 없으나 쓰기가 불편하다. 같은 음소를 적는 글자가 둘 이상일 때 어느 글자를 써야 할 지, 일일이 기억해 두지 않으면 바르게 쓸 수가 없다. 영어의 [k] 음소는 'k' 자로도 때로는 'c'로도 적는데, 어떤 경우에 'k' 자를 쓰고 어떤 경우에 'c' 자를 쓰는가는 단어에 따라 달리 쓰인다고 말할 수밖에 없다. 반면에 음소 수보다 글자 수가 적으면 쓰기는 쉬우나 읽기가 불편하다. 한 글자가 둘 이상의 음소를 표기하는데 쓰인다면 그 언어에 능통한 사람이 아니고서는 어떻게 읽어야 할 지 알기 힘들기 때문이다.[3]
한국어는 한 음절이 하나의 뜻덩이[형태소나 단어]를 나타내는 경우가 많이 있기 때문에 이와 같은 한국어의 특징을 살려 뜻을 이해하기 쉽게 하려면 글자를 음절 단위로 모아 쓰는 것이다. 한글 맞춤법은 우리말의 이와 같은 특징을 살려 표기하도록 음절 단위로 모아 쓸 뿐만 아니라 각 형태를 고정시켜 표기하도록 규정하고 이다.
한글을 전산화 하여 데이터로써 처리하기 위하여 다양한 한글 문자셋이 존재하고 있다.
초성, 중성, 종성을 분리하여 저장하는 조합형 한글 문자셋과 하나의 음절을 2bytes로 저장하는 완성형 한글 셋이 사용되고 있으나, 저장공간의 효율성을 고려하여 산업계에서는 KSC5601, CP949등의 완성형 한글셋을 널리 사용하고 있다.
또한 한글의 음절을 하나의 문자로 사상시킴으로써 음소단위로 사상되어진 Roman 문자와 데이터베이스상에 데이터의 정정을 위해서는 음소단위의 1단계 분해작업 이후에 현재까지 알려진 데이터의 정정기법 등을 적용 가능한 상태에 이를 수 있게 된다.
한글 데이터 내에 존재하는 오류 유형을 분리하여 어떠한 유형의 오류가 존재하는지를 파악하기 위해 국내 모 생명보험사의 고객데이터 내에 존재하는 오류를 샘플링하여 다음과 같이 오류 유형으로 구분하였다.

오류 유형		오류의 예
하나의 음절(문자)분리		"가"라는 하나의 음절이 "ㄱ" "ㅏ"로 표기됨(2code 사용)
모음 오류	Typing 오류	"논현구"를 "논현구"로 입력
	Shift+ Typing오류	"대방동"을 "대방동"으로 Shift Key 입력
자음 오류	Typing오류	"경기도"를 "령기도"등으로 입력
	Shift+ Typing오류	"서울시"를 "써울시"로 입력
유사 문자 오류		"논현동"을 "론현동"으로 입력
영문자로 한글 Typing		"서울시"를 "tjdnftl"로 입력
복합적 오류		상기 예시된 오류가 복합적으로 발생
인식 불가		사람의 수작업을 통하여도 인식 불가능

표 1 한글 데이터의 오류 분류

한글 입력시 현재 사용하고 있는 2벌식 자판배열 한글의 입력 속도를 높이는 반면 모음 및 자음 등이 유사 음소끼리 묶여있어 데이터 입력에 의한 오류가 많은 편이다.[4]

4. 한글데이터의 오류정정을 위한 프로세스 구조

데이터의 오류를 정정하기 위해서는 데이터의 오류 수준을 측정하기 위한 검사 / 오류 측정 / 데이터 정정의 프로세스를 구성하게 되는데 한글의 오류 수준을 검사를 위해 제안되는 음운 유사성 검사와 자판기 타이핑 입력 오류검사를 사용하여 유사성을 검사하고, 데이터의 유사도가 음절길이에 따라 조정된 임계치를 만족할 경우 데이터의 정형화를 실시하게 된다.

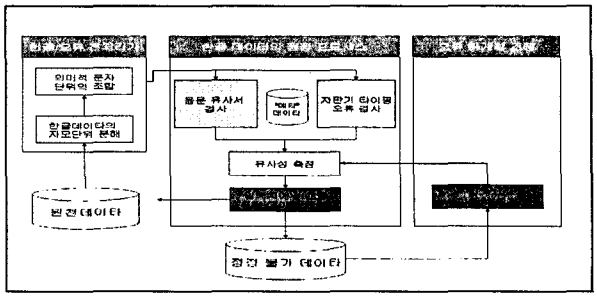

그림 1 한글데이터의 정정 프로세스 구성도

이번 논문을 통하여 한글데이터의 정정 프로세스와 오류 수준 조정 프로세스에 대하여 다루고 있다.

4.1 음운 유사성 검사 프로세스
자모단위의 유사성 검사를 위해서는 음운 유사성(Phonetic Similarity)에 비교하여 평가하는 방법이 사용되고 있다.[2]
음운이 비슷한 음소끼리 그룹화하여 음절상에 나타나는 첫번째에 자음은 음소문자를 그대로 사용하고 하나의 음절에서 이후의 자음은 수치에 매핑하여 유사도를 검사하며 이때 모음은 무시하여 유사성을 측정한다[3]

4.2 자판기 타이핑 오류 검사 프로세스
우리가 현재 일반적으로 사용하고 있는 자판은 영문자판의 배열 위에 코드를 Shift하여 한글을 입력하도록 되어 있어 입력하고자 하는 음소를 입력 시 자판 배열상의 유사키를 입력하는 오류가 발생하게 된다.[4]
음소 단위 문자입력 시 자음을 입력하고자 하는 경우 오

타로 인하여 모음 입력 시 발견할 확률이 높으므로 이를 오류 유형에서 분리하였다.

4.3 이질성 함수를 이용한 유사도 측정

유사성을 검사하기 위해서는 이질성 함수(Difference function) d(x,y)를 사용하는데, 이때 0에서 1사이의 값을 가지게 되며 0은 완전 일치(Exact match)를 의미한다. 다수의 속성에 대해 유사성을 검사 시는 중요도에 따라 가중치를 두어 유사성을 두어, 2개의 벡터 x0,…,xn과 y0,…,yn의 이질성을 구하기 위해 사용한다[2]

$$\frac{\sum_{i=0}^{n} w_i \bullet d(x_i, y_i)}{\sum_{i=0}^{n} w_i}$$

이번 논문을 통하여 제안하는 데이터베이스상의 데이터 정정 기법의 경우 소리 데이터의 음운의 유사성이 아닌 2bytes로 코드화되어 있는 Text형태의 데이터 정정에 목적을 두고 있다. 이를 위하여 음운 유사성과 자판기 사용에 의해 발생하는 오류에 대한 2가지 벡터를 가지고 유사성 검사를 사용하고자 한다. 이때 두 가지 항목의 가중치는 샘플링 된 데이터에 적합성을 검토하여 가중치를 조정하여 최적의 값을 찾고자 한다.

4.4 오류 수준의 측정 프로세스

데이터베이스상의 오류 데이터의 정정을 위해서 사용되는 이질성 함수 적용 시 어느 수준의 이질성까지를 유효한 데이터로 취급할 것인가 하는 임계치를 정하여야 하는 문제가 있다.

하나의 항목이 하나의 음절로 구성되어 있고 초성과 중성으로만 구성된 하나의 음소가 일치하지 않는 경우 이질성이 0.5의 값을 같게 되나 3개의 음절로 구성되어 있고 초성 및 중성으로 가정한다면 하나의 음소가 틀린 경우 0.167의 이질성을 가지게 된다. 따라서 음절의 길이에 따라 다른 일치성 판단을 위한 다른 임계값을 가지는 것이 보다 일치성 판단의 효율성을 높일 수 있다. 다수의 음절로 구성된 경우 보다 낮은 이질성 평가값을 수용할 수 있다.

5. 자모단위 정정기법의 성능 측정

제안된 자모단위 비교를 통한 한글데이타 정정 기법의 평가를 위하여 고객정보에 대해 정합성이 문제가 있는 300 개를 샘플링하여 제안되어진 자모단위 비교를 통한 정정기법을 음절단위 정정기법과 성능을 비교하였다.

적용 기법	정정률
음절단위 비교 정정 기법	21.0%
제안되어진 음소단위 비교 정정	24.7%

표 2 제안 기법의 성능 평가

제안되어진 음소단위 비교 정정기법은 3.7% 정정률이 높았다. 그 외 정정이 어려운 데이터로는 무의미한 코드 또는 값을 입력한 경우가 52% 존재하였다.

6. 결론 및 향후 연구 방안

데이터 정형화를 통하여 얻을 수 있는 이득으로는 데이터의 통합 및 Aggregation시 데이터를 보다 정확한 형태로 모아 낼 수 있고, 오류 유형에 대한 Account-Ability에 대한 향상이 가능하게 되는 것이다.

제시되는 방법을 통해 유사 음운에 대한 정형화뿐 아니라 한글 입력 시 발생하는 오류를 정정하므로 서 정정률의 향상을 얻을 수 있다.

데이터베이스상에 존재하는 한글 데이터의 경우 단순히 유사 음운에 대한 비교뿐 아니라 데이터의 입력시점에 사용되는 자판기를 통한한 Interface에 의한 오류를 감안하여 한글 데이터의 정정시 보다 효율적인 데이터의 오류를 감지해 낼 수 있게 된다.

본 논문에서는 이러한 데이터입력의 환경을 고려하여 데이터를 정정할 수 있는 기법을 제시하였다. 향후 제시된 방법을 통하여 한글 도메인 값의 정정을 통한 데이터의 표준화를 통하여 Data Warehouse 및 운영CRM 구축 시 활용하고자 한다. 길이가 다른 음절의 동일 도메인 안의 데이터의 적정 수준의 임계치를 찾아 적용업무에 맞게 조정하는 방법에 대하여 향후 연구를 진행하여야 한다.

참고

[1] Larry P. English, "Improving Data Warehouse and Business Information Quality", WILEY, 1999

[2] David Loshin, "Enterprise Knowledge Management - The Data Quality Approach", Morgan Kaufmann, 2001

[3] 김영제, "조합형 문자 분석을 이용한 한글 문서 인식 알고리즘 개발연구",서울시립대학교 대학원, 1998

[4] 유택상, "컴퓨터 한글자판에서 연타가 타자속도에 미치는 영향에 관한 연구", 한양대 대학원, 1994

[5] Michael H. Brackett, "The DataWarehouse Challenge-Taming Data Chaos", WILEY 1996

[6] Ross, Ronald G., "Data Dictionaries and Data Administration : Concepts and Practices for Data Resource Management.", McGrawHill 1984

[7] Thomas Flanagan, "A Practical Guide to Archiving Enterprise Data Quality", Techguide Ltd, 1998

[8] "Trillium Software WhitePater","Total Global Data Quality",Harte-Hanks, 2002