건설분야 텍스트 인식시스템의 매칭알고리즘 개발

송종관*, 정 숙*
*한국건설기술연구원
e-mail: {song5216, white1509}@kict.re.kr

Development of Matching Algorism for System Recognizing Text in the Construction Field

Jong-Kwan Song*, Suk Jeong*
*Korea Institute of Construction Technology

요 약

현재 모든 분야에 IT산업이 융합되어 있지만 건설분야에서는 IT산업과의 융합이 많이 시도되고 있음에도 불구하고 타 산업에 비해 미비한 실정이다. 특히, 설계단계 공사비정보는 설계자의 의사결정을 지원하는 중요한 자료원임에도 불구하고 작성자에 따라 내역서에 쓰이는 작업항목 및 규격의 표현방식이 다르고 외래어 표음 및 오타, 그리고 부가정보 표기로 인해 단가축적의 시스템 및 DB화 가 이루어지지 않고 있다. 따라서 본 연구는 시공단계에서 발생된 실적단가를 설계단계에서 효과적으로 활용하기 위해 동일한 작업항목의 상이한 표현을 동일하게 인식할 수 있는 텍스트 인식시스템의 알고리즘을 제시한다. 텍스트 인식알고리즘에는 “유사어 및 단어테이블”, “기준작업항목 테이블”, “인식된 작업항목 테이블” 등으로 구성된 DB, 최소의미단위 단어를 비교·분절하기 위한 문자열 매칭 알고리즘, 그리고 동일하지 않은 텍스트를 인식하고 사용자의 의사결정을 지원하기 위한 유사도 계산으로 구성하였다.

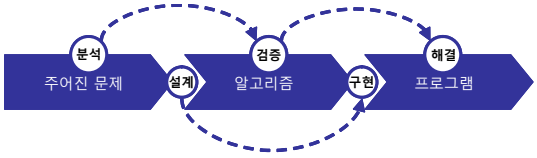
1. 서론

정부는 2000년 이후 건설분야 정보화를 위해 건설정보화 사업을 진행중에 있으며, 건설CALS 등의 정보화 사업을 통해 그 목적을 이루어 가고 있다. 하지만 건설산업은 노동력 중심의 1차 산업부터 첨단 공법이 어우러지는 포괄적인 산업형태의 특성을 가지고 있으며, 주문제작형 산업이기 때문에 건설공사의 표준화 및 시스템화를 이루는데 타 산업에 비해 많은 어려움을 가지고 있는 실정이다. 또한 기획설계단계부터 유지관리단계까지 업무를 처리하는 주체가 다르기 때문에 각 단계별로 발생되는 정보의 공유 및 시스템화가 이루어지지 못하고 있는 실정이다.
실적공사비는 한국건설기술연구원에서 1년에 2회 단가정보를 갱신하여 고시하는 『건설공사 실적공사비 적용 공종 및 단가』와 기존 프로젝트의 실행내역서에서 그 정보를 획득할 수 있다. 그러나 『건설공사 실적공사비 적용 공종 및 단가』또한 실행내역서에서 각 공종별 단가를 획득하기 때문에 시공단계 발생된 실행내역서의 단가를 축적 및 활용하는 것은 매우 중요하다. 하지만 내역서에 쓰이는 작업항목 및 규격을 내역서 작성자에 따라 표현방식이 다르고, 외래어, 오타, 및 부가정보표현으로 인한 불일치 등의 문제들로 인해 실적공사비를 DB화 하는데 상당한 어려움이 있는 실정이다.[1]

따라서 본 연구는 시공단계 발생되는 실적단가를 설계단계에서 효과적으로 활용하기 위해 동일한 작업항목의 상이한 표현을 동일하게 인식할 수 있는 텍스트 인식 알고리즘을 제시하고자 한다.

2. 알고리즘

알고리즘(Algorithm)은 범용 환경 또는 특정분야에서 주어진 문제를 해결하기 위해 이를 어떤 구조와 방법으로 구성해 갈 것인지를 생각하는 기본적인 설계에 해당한다. 예를 들어, 컴퓨터의 처리과정을 살펴보면, 문제의 대상이 되는 데이터를 입력하고 이를 바탕으로 원하는 출력물을 만들어내는 과정을 수행하게 된다. 이때 컴퓨터로 문제를 해결하는 것은 특정 프로그래밍언어를 사용하더라도 프로그래밍언어들이 허용하는 명령어들의 표현 한계 내에서 구현할 수 있도록 논리 정연하게 계획하는 것을 의미한다. 이처럼 주어진 문제의 해결을 위해 논리 정연한 절차에 의해 계획한 해결방안을 정형적이고 체계적으로 기술한 것을 알고리즘 이라고 한다. 그림 1은 알고리즘을 활용한 문제해결 절차를 보여주고 있다.[3]



<그림 1> 알고리즘을 활용한 문제해결 절차

알고리즘은 어떤 문제를 해결하기 위한 개산적인 문제 해결 절차이며, 이 절차를 통해 요구하는 결과가 도출된다. 알고리즘에는 여러 문제들의 문자 또는 숫자들이 입력되고 알고리즘이 실행된 후 문제에 대한 결과를 출력할 수 있다. 문제를 해결하고 설계된 알고리즘을 토대로 작성될 프로그램을 제대로 사용하기 위해 다음 표 2와 같은 몇 가지의 요건을 만족해야 한다.

<표 2> 알고리즘의 요건[3]

요 건	내 용
정확성	알고리즘은 입력을 받아들이고 이것을 토대로 결과물을 만들어냄
유한성	알고리즘은 일정한 단계가 실행된 후에는 반드시 종료되어야 함
명확성	알고리즘이 명확하다는 것은 모호하지 않고 이해하기 쉬움을 의미함
유효성	알고리즘에서 요구하는 연산은 이산적인 컴퓨터에서 처리될 수 있는 계산이어야 함

2.1 텍스트 인식시스템 필요성

1)실적단가 축적의 어려움

실적공사비의 대상 범위는 공공공사의 계약단가를 기준으로 산정하게 법에서 정하고 있으며, 내역서의 신뢰도 관점에서 공공기관에 의해 확인된 공공공사의 입찰내역서가 그 대상이 된다. 현재 이러한 단가를 축적하고 시스템화 하기위한 인프라가 구축되어 있지 않기 때문에 이에 대한 연구 예산이 확보되지 않은 실정이다. 또한 작업항목의 명칭이 통일되어 있지 않고 작성자에 따라 다르게 표현되어 있기 때문에 단가 축적을 시스템화 하기에는 어려움이 있는 것으로 조사되었다.

2)동일항목의 다른표현

내역서의 품명은 동일 항목의 경우 대체적으로 동일하게 표현된다. 하지만 “방수몰탈”과 “방수모르타르바름”, “방수몰탈바름”은 동일한 항목이지만 내역서 작성자에 따라 다르게 표기된 예이다. 이처럼 동일품목의 다른 표현이 내역서 내에 상당부분 존재하는 것으로 조사되었다. 이것은 실적단가를 축적하는 기준으로 사용되는 품명에서 동일항목의 다른 표현을 동일하게 인식하게 함으로써 시스템화를 할 수 있을 것이다.

3)외래어 표현

작업항목을 조사한 결과 외래어의 국어식 표현으로 인한 표기의 상이함이 조사되었다. 이는 표음문자인 국어의 특성이 반영된 것으로써, 예를 들어 국어식 표현인 “문”은 “도어”, “도아” 등으로 표기되고 있는 것으로 조사되었다. 또한, “방수모르타르바름”과 “방수몰탈바름”에서 “모르타르”와 “몰탈”은 동일 의미의 다른 표현으로 표기되고 있다.

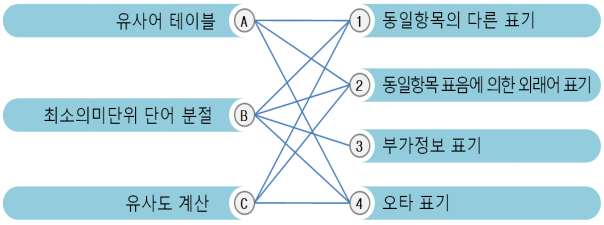
국어와 외국어의 병행사용에 대한 문제 또한 조사되었다. "EXP. JOINT"는 “신축줄눈”과 동일한 작업항목이지만 외국어의 약어를 사용한 예이다. 이러한 대표적인 예로써 콘크리트는 한글표현인 “콘크리트”와 영어표현인 “Concrete”, 그리고 약어인 “Con′c”를 내역서 작성자의 선호에 따라 사용되고 있는 것으로 조사되었다.

이러한 내역서 표기방식의 문제는 유사어 테이블을 구축하여 해당 품목의 다른 표기를 축적함으로써 동일한 항목으로 인식하게 할 수 있을 것이다.

3. 텍스트 인식알고리즘

3.1 텍스트 인식알고리즘의 개발방향

실적단가를 축적하기 위한 목적으로 내역서 작업항목 품명의 명칭을 동일하게 인식하게 하는 시스템을 개발하기 위해 해결해야할 사항은 ①동일항목의 다른표기, ②동일항목의 표음에 의한 외래어 표기, ③부가정보 표기, ④오타 표기문제이다. 이러한 문제는 다음 그림 2와 같이 ①, ②, ④항목은 Ⓐ유사어 테이블을 이용하여 동일의미의 여러 표현을 동일하게 인식하게 하고, ①, ②, ③, ④항목은 Ⓑ최소의미 단위 단어의 분절을 통해 부위 및 공법정보를 인식하게 하며, 이는 문자열 매칭알고리즘을 활용하여 단어분절을 실시한다. ①, ②, ④항목의 단어별 다른 표기는 Ⓒ유사도계산을 통해 유사한 항목을 사용자가 선정할 수 있도록 지원함으로써 인식하게 한다.



<그림 2> 인식시스템 개발 방향

3.2 텍스트 매칭 알고리즘 선정

작업항목 단어를 분절하기 위해 비교를 수행할 적절한 매칭알고리즘이 필요하다. 문자열 매칭알고리즘을 선정하기 위해 위의 절에서 원시적 매칭알고리즘, 오토마타를 이용한 매칭알고리즘, 라빈-카프 매칭 알고리즘, KMP 매칭 알고리즘, 보이어-무어-호스풀 알고리즘에 대하여 고찰하였다. 선정기준으로는 알고리즘의 수행시간과 문자열 매칭 알고리즘의 시간을 측정하고 작업항목 품명에 적합한 짧은 문자열에 알맞은 알고리즘을 선정하여야 한다. 이를 위해 각 문자열 알고리즘의 특성과 시간을 비교하였다. 비교를 수행하기 위해 건축공사의 단어표현이 비교적 다양한 방수공사를 선정하여 비교를 수행하였다. 각각의 문자열 매칭 알고리즘을 시스템으로 구현하여 동일한 내역서를 실행한 결과 표 3과 같이 5개의 매칭알고리즘 모두 41.67ms의 미세한 시간이 측정되었다. 이는 시스템 처리속도를 측정할 만큼의 데이터를 처리하지 않았기 때문으로 사료되며, 시스템 구동사양의 최소시간으로 측정되었다. 건축 도메인의 특성상 내역서 작업항목 품명은 단어의 조합으로 여러 작업항목이 구성되기 때문에 처리속도에

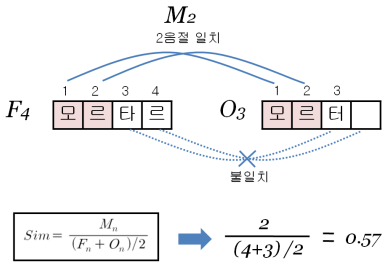
크게 영향을 받지 않을 것으로 사료되며, 문자열 비교방식에 의한 특성으로 적정문자열을 선정하였다. 따라서 작업항목 품명을 비교하는 짧은 문자열이기 때문에 이론상의 처리속도가 $\Theta(\frac{n}{m})$만큼 빠르고 패턴의 위치에 대해 불일치 시 점프를 통해 이동하며 매칭을 수행하는 보이어 무어 호스풀 알고리즘이 작업항목 인식을 위해 적합할 것으로 사료된다.

<표 3> 문자열 매칭 알고리즘의 비교

매칭방식	현황
전통적 매칭	처리시간 : $O(nm)$
	실제 문자열 처리시간 : 41.67ms
	특성 :순차적으로 모든 문자를 비교
오토마타 매칭	처리시간 : $\Theta(n+\lvert\sum\rvert m)$
	실제 문자열 처리시간 : 41.67ms
	특성 : 매칭 상태에 따라 문자를 찾아가는 방식
라빈_카프 매칭	처리시간 : $\Theta(n)$
	실제 문자열 처리시간 : 41.67ms
	특성 : 문자열 패턴을 수치화 하여 수치비교로 매칭을 수행
KMP 매칭	처리시간 : $\Theta(n)$
	실제 문자열 처리시간 : 41.67ms
	특성 :패턴의각 위치에 대해 매칭에 실패 시 돌아갈 곳을 알려줌
보이어_부어_호스풀 매칭	처리시간 : $\Theta(\frac{n}{m})$
	실제 문자열 처리시간 : 41.67ms
	특성 : 패턴의 위치에 대해 불일치 시 점프를 통해 이동함

3.3 유사도 계산

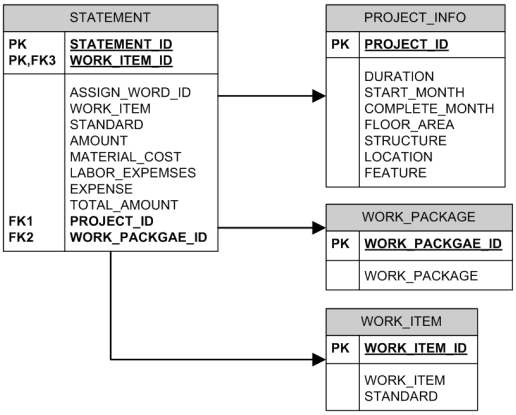
텍스트를 인식하기 위해 문자열 매칭 알고리즘을 통해 검색을 수행한 후 검색되지 않은 단어에 한하여 유사도계산을 수행한다. 단어와 단어의 음절, 작업항목 품명에 할당된 식별자와 기준작업항목 테이블의 품명에 할당된 식별자간의 유사도를 계산하기 위해 Oliver에 의한 유사도계산식을 사용하였다. 이 계산식은 PHP 프로그래밍언어의 Similar_Text 함수에 활용되고 있으며, 두 단어간 유사도 계산을 수행한다.[5](그림 3참조)



<그림 3> 유사도계산 예시

3.4 인식된 작업항목 테이블

인식과정을 통해 인식된 작업항목의 데이터를 저장할 목적으로 그림 4와 같이 인식된 작업항목 테이블을 구성하였다. 이 테이블은 향후 데이터의 효율적인 활용을 위해 공사위치, 공사시점, 건물구조 등의 개요정보를 속성으로 포함하고 있다. 내역서(STAETMENT)테이블은 프로젝트정보(PROJECT_INFO)테이블의 개요정보, 공종(WORK_PACKAGE)테이블의 공종정보, 그리고 기준작업항목테이블에 사용된 작업항목(WORK_ITEM)테이블의 작업항목 품명정보를 포함하고 있다. 그리고 내역서식별자(STATEMENT_ID)와 작업항목식별자(WORK_ITEM_ID)를 PRIMARY KEY로 내역서의 데이터항목들을 속성으로 정의하였다.



<그림 4> 인식된 작업항목 테이블의 EF Diagram

4. 결 론

본 연구는 텍스트 매칭 알고리즘, 단어간 유사도 계산, 그리고 작업항목 테이블을 활용한 DB를 통해 동일의미의 다른 표현을 수집함으로써 동일한 의미의 다른 표현을 동일한 작업항목으로 인식시킬 수 있었다. 텍스트 매칭 알고리즘은 처리속도 및 특성을 파악하여 가장 적합한 매칭알고리즘을 선정하였으며, 유사도계산은 음절간의 동일한 정도를 수치화하여 유사도를 계산하는 방식으로 2바이트인 한글의 유사도 평가에 적합할 것으로 사료되며, 동일한 의미의 다른 표현을 수집함으로써 지속적인 정확도 향상을 이룩할 수 있을 것으로 사료된다.

참고문헌

[1] 구교진, 송종관, 박성철, 박성호. 2008, “실적단가 활용을 위한 작업항목 매칭 프로세스 모델”, 대한건축학회, 대한건축학회논문집, 제24권 제6호

[2] 문병로. "쉽게 배우는 알고리즘-관계중심의 사고법",한빛미디어, 2007

[3] 박지연 (2007), 알고리즘의 이해, 기한재

[4] 송종관. 2008, “실적공사비 기반 견적시스템의 단가관리를 위한 작업항목 매칭 프로세스”, 석사학위논문, 서울시립대학교

[5] PHP 매뉴얼, http://www.php.net/similar_text