이질적 색인어의 가중치 합에 기반한 수식 검색 시스템

(An Equation Retrieval System Based on Weighted Sum of Heterogenous Indexing Terms)

신 준 수 † 김 학 수 ††
(Junsoo Shin) (Harksoo Kim)

요 약 다양한 수식을 포함하는 수학 문서들을 효과적으로 검색하기 위해서는 수식 인지 검색 엔진이 필요하다. 본 논문에서는 구조적으로 유사한 수식들을 효과적으로 찾아주는 수식 검색 시스템을 제안한다. 제안 시스템은 MathML 수식들을 연산자, 변수, 그리고 수식 구조와 같은 3가지 형태의 이질적 색인어로 분리하고 독립적으로 색인한다. 사용자가 MathML 수식을 입력하면 제안 시스템은 이질적인 색인어들을 위한 3가지 언어모델들의 가중치 합을 이용하여 수식들을 검색하고 순위화한다. 244,824개의 MathML 수식을 대상으로 한 실험에서 제안 시스템은 비공개 테스트에서 53%의 1순위 정확률, 공개 테스트에서 63%의 1순위 정확률을 보였다.

키워드 : 수식 검색, 이질적 색인어, MathML

Abstract To effectively retrieve mathematical documents including various equations, math-aware search engines are needed. In this paper, we propose a equation retrieval system which helps users effectively search structurally similar equations. The proposed system disassembles MathML equations into three types of heterogeneous indexing terms; operators, variables, and partial structures of equations. Then, it independently indexes the disassembled terms. When a user inputs a MathML equation, the proposed system searches and ranks equations using weighted sums of three language models for the heterogeneous indexing terms. In the experiments with 244,744 MathML equations, thre proposed system showed reliable performances (a P@1 of 53% in the closed test and a P@1 of 63% in the open test).

Key words : Equation retrieval, Heterogenous indexing term, MathML

1. 서 론

인터넷의 대중화와 함께 Web2.0시대가 도래함에 따라 웹에서 생산되는 문서의 양이 계속해서 증가하고 있다. 그에 따라 논문, 특허, 위키피디아(wikipedia) 등 수식이 들어간 기술 문서들도 점차 늘어나고 있는 추세이며, MathML(Mathematical Markup Language)[1]이 제정된 이후로는 MathML 수식을 포함하는 기술 문서들이 꾸준히 증가하고 있다. 그러나 기존의 텍스트 검색 시스템을 사용하여 기술 문서 내의 특정 수식을 찾는 것은 거의 불가능하다. 그러므로 수식을 효과적으로 검색하기 위해서는 이미지 검색, 음성 검색과 같이 수식 검색이라는 특수한 목적을 가진 검색 시스템이 필요하다. 기존의 수식 검색 시스템은 색인을 하기 위한 비용이 필요하거나, 수학식의 구조를 반영하지 못했다는 단점이 있다. 본 논문에서는 MathML 기반의 수식을 효과적으로 찾아주는 새로운 검색 시스템을 제안한다. 제안 시스템은 수식으로부터 알아낼 수 있는 연산자 정보, 변수 정보, 그리고 구조 정보를 독립적으로 색인하고 각각에 독립된 가

· 이 논문은 2008년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(KRF-2008-313-D00907)

† 학생회원 : 강원대학교 컴퓨터정보통신공학과
nlpsjs@kangwon.ac.kr
†† 정 회 원 : 강원대학교 컴퓨터정보통신공학과 교수
nlpdrkim@kangwon.ac.kr

논문접수 : 2010년 4월 23일
심사완료 : 2010년 8월 18일

정보과학회논문지: 소프트웨어 및 응용 제37권 제10호(2010.10)

중치를 부여하여 수식을 순위화함으로써 완전히 일치하지 않는 수식도 검색이 가능하다는 장점이 있다.

본 논문의 구성은 다음과 같다. 2장에서 수식 검색과 관련된 기존의 연구에 대해서 살펴보고, MathML에 대해서 설명한다. 3장에서는 제안한 수식 검색 시스템의 구조를 상세하게 기술한다. 4장에서는 제안 시스템의 유용성을 평가하기 위한 몇 가지 실험을 진행하고, 5장에서 결론을 맺는다.

2. 관련연구

수식 검색과 관련된 기존 연구는 수식으로부터 색인어를 효과적으로 추출하는 방법에 대한 연구[2,3]와 수식 검색 결과를 순위화하는 방법에 대한 연구[4]로 나눌 수 있다. 색인어 추출과 관련된 대표적인 연구는 정규 표현을 이용하여 수식의 색인어를 생성하는 연구[2]와 후위표기법을 이용하여 색인어를 추출하는 연구[3]가 있다. Adeel 외[2]는 MathML 태그로 구성된 정규 표현을 구축한 후, 각 정규 표현에 소수, 지수, 집합, 행렬 등 특정 키워드(keyword)를 부여하였다. 그리고 수식을 구축된 정규 표현과 비교한 후, 매칭(matching)된 수식의 키워드를 색인하였다. 그러나 Adeel 외의 연구는 정규 표현과 키워드를 일일이 수동으로 구축해야 한다는 문제점이 있다. Misutka와 Galambos[3]는 수식을 확장하여 유사한 의미를 갖는 n개의 수식을 생성하고, 생성된 수식으로부터 후위표기법을 이용하여 색인어를 추출하는 방법을 제안하였다. Misutka와 Galambos의 연구는 유사 수식 검색이 가능하여 재현율을 높일 수 있다는 장점이 있지만 정확률이 떨어지고 순위화 방법도 명확하지 않다는 단점이 있다. 수식 순위화와 관련된 대표적인 연구로는 함수명, 변수, 상수에 고정 가중치를 주는 연구[4]가 있다. Youssef[4]는 'sin', 'tan', 'cos'과 같은 함수명에 가장 높은 가중치를, 연산자에 두 번째로 높은 가중치를, 변수와 상수에 가장 낮은 가중치를 부여하였다. 이는 질의의 형태와는 상관없이 일정한 가중치를 부여함으로써 수식의 구조를 반영하지 못하는 단점이 있다. 상기한 문제들을 해결하기 위해서 본 논문에서는 수식을 연산자, 변수, 구조로 분류하여 독립적으로 색인하고, 각각에 가중치를 할당하는 새로운 수식 검색 시스템을 제안한다. 본 논문에서는 검색의 대상이 되는 수식들이 그림 1과 같이 구조적 정보를 포함하는 MathML로 기술되어 있다고 가정한다. 그림 1에서 '<mi>', '<mo>', '<mn>', '<mfenced>', '<msup>'은 변수, 연산자, 상수, 괄호, 지수를 의미한다.

화면 표시	내부 표현
$(a+b)^2$	<mrow> <msup> <mfenced> <mrow> <mi>a</mi> <mo>+ </mo> <mi>b</mi> </mrow> </mfenced> <mn>2</mn> </msup> </mrow>

그림 1 MathML의 예

3. 수식 검색 시스템

본 논문에서 제안하는 수식 검색 시스템의 구조는 그림 2와 같다. 그림 2에서 보는 것과 같이 제안 시스템은 색인 시스템과 검색 시스템으로 구성된다.

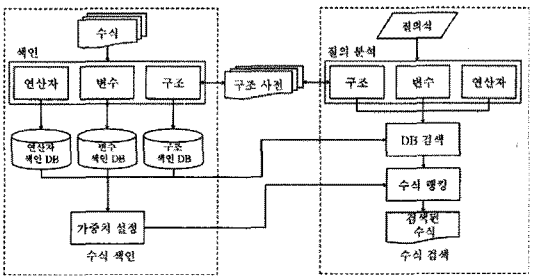

그림 2 수식 검색 시스템 구조도

색인 시스템은 MathML 수식을 분해하여 분석한 후, 변수 정보, 연산자 정보, 구조 정보를 독립적으로 색인한다. 이 때 분수, 행렬, 밑수, 지수 등과 같은 수식 구조는 MathML의 특정 태그를 활용한다. 검색 시스템은 MathML 질의를 분석한 후, 변수 정보, 연산자 정보, 구조 정보에 서로 다른 가중치를 부여하여 순위화한다.

3.1 수식 색인어 추출 및 저장

대부분의 간단한 수식은 변수와 연산자로 이루어져있다. 그러나 논문이나 특서, 연구보고서와 같은 기술 특화된 문서에는 분수, 밑수, 지수 등과 같이 복잡한 구조의 수식이 자주 출현한다. 이런 복잡한 수식은 변수와 연산자만으로 그 의미를 표현하는데 한계가 있다. 예를 들어, 표 1과 같은 두 개의 수식이 있다고 가정하자. 두 수식에서 변수와 연산자만을 색인어로 추출한다면 '예시 1'과 같이 두 수식은 전혀 다른 의미를 가지지만 동일한 색인어를 갖게 된다. 이러한 문제점을 해소하기 위해서 제안 시스템은 '예시 2'와 같이 변수, 연산자와 함께 수식 구조도 색인어로 추출한다.

수식 구조 색인어는 표 2와 같은 구조 사전을 이용하여 추출하였으며, 실제로 수식의 의미 및 구조를 나타내는 것만을 대상으로 하였다.

표 1 색인어 추출의 예

수식	예시 1		예시 2	
$(a+b)^x$	변수	a, b, x	변수	a, b, x
			연산자	+
	연산자	+	구조	mi mo mfenced msup
$ax+b$	변수	a, b, x	변수	a, b, x
			연산자	+
	연산자	+	구조	mi mo

표 2 구조 사전

태그	구조 예	태그	구조 예
mtable	행렬	mover	$\overrightarrow{A}$
mtr	행렬	munder	$\lim_{a\to 0}$
mtd	행렬	munderover	$\sum_{a=0}^{n}$
msqrt	$\sqrt[a]{A}$	mfenced	$(A+B)$
mroot	$\sqrt{A}$	mfrac	$\frac{A}{B}$
msup	A^2	mo	연산자
msub	A_2	mn	상수
msubup	$\int_a^b$	mi	변수

표 2에서 보듯이 구조 사전은 16개의 태그(tag)로 구성되어 있다. 표 2는 각 태그가 나타내는 구조의 예를 보여준다.

수식 색인어 추출이 끝나면 각 색인어를 독립적으로 역파일 형태의 데이터베이스에 저장한다. 각 색인어를 독립적으로 저장하는 이유는 다음과 같다. 표 3과 같은 두 개의 색인식(indexed equation; '색인식-1'과 '색인식-2')과 한 개의 질의식(queried equation; '질의식-1')이 있다고 가정하자. 각 색인어를 독립적으로 색인하지 않는 방법에서는 '질의식-1'과 '색인식-2'의 모든 색인어가 'msup'을 제외하고는 동일하다. 그러나 '질의식-1'과 의미적으로 유사한 '색인식-1'는 두 개의 색인어('c'와 'd')를 포함하지 않는다. 이처럼 각 색인어를 독립적으로 색

표 3 일반 수식 색인 예제

	수식	색인어
질의식-1	$c+d^2$	c, +, d, 2, mi, mo, mn, msup
색인식-1	$a+b^2$	a, +, b, 2, mi, mo, mn, msup
색인식-2	$2c+d$	c, +, d, 2, mi, mo, mn

표 4 독립적 수식 색인 예제

	수식	색인어		
		구조	연산자	변수
질의식-1	$c+d^2$	mi, mo, mn, msup	+	c, d
색인식-1	$a+b^2$	mi, mo, mn, msup	+	a, b
색인식-2	$2c+d$	mi, mo, mn	+	c, d

인하지 않으면 동일한 의미지만 변수가 다른 수식들이 낮은 순위로 밀리는 문제가 발생할 수 있다.

이러한 문제를 해결하기 위해서 제안 시스템은 연산자, 변수, 구조에 가중치를 부여할 수 있도록 표 4에서 보는 것과 같이 독립적인 색인을 수행한다.

3.2 수식 검색 및 순위화

제안 시스템은 식 (1)과 같이 색인식 IE에서 질의식 QE가 생성될 확률을 의미하는 언어모델[5]을 이용하여 수식을 순위화한다.

$$P(QE|IE) = \prod_{q \in QE} P(q|IE) \tag{1}$$

연산자, 변수, 구조에 독립적으로 가중치를 부여하기 위해서 식 (1)을 식 (2)와 같이 확장한다.

$$P(QE|IE) = \alpha \prod_{s \in QEs}^{|s|} P(s|IEs) + \beta \prod_{o \in QEo}^{|o|} P(o|IEo) + \gamma \prod_{v \in QEv}^{|v|} P(v|IEv) \tag{2}$$

식 (2)에서 s는 수식 구조를 나타내는 MathML 태그를 의미하고, $|s|$는 질의식내 수식 구조를 나타내는 태그의 수를 의미한다. o는 연산자를 의미하고, $|o|$는 질의식내 연산자의 수를 의미한다. v는 변수를 의미하고, $|v|$는 질의식내 변수의 수를 의미한다. IEs, IEo, IEv는 각각 색인식 IE의 수식 구조, 연산자, 변수를 의미한다. α, β, γ는 각각 구조, 연산자, 변수에 대한 가중치 요소로써 식 (3)과 같이 결정한다. $|IEs|$, $|IEo|$, $|IEv|$는 각각 색인된 전체 수식에서 구조를 나타내는 태그의 수, 연산자의 수, 변수의 수를 나타낸다. 이는 시스템에 색인된 전체 수식에 따라 서로 다른 가중치 값을 갖게 된다. 본 논문에서 결정된 가중치 요소 α, β, γ는 각각 0.61, 0.18, 0.21의 값을 갖는다.

$$\alpha = \frac{|IEs|}{|IEs|+|IEo|+|IEv|}$$
$$\beta = \frac{|IEo|}{|IEs|+|IEo|+|IEv|}$$
$$\gamma = \frac{|IEv|}{|IEs|+|IEo|+|IEv|} \tag{3}$$

4. 실험 및 평가

검색 시스템을 올바르게 평가하기 위해서는 공인된 테스트 컬렉션(test collection)이 필수적이다. 그러나 수

식 검색 시스템을 평가하기 위해 구축되어 있는 테스트 컬렉션은 찾을 수 없었으며, 아직 존재하지 않는 것으로 보인다. 그래서 본 논문에서는 arXMLiv[6]의 논문 500개에 포함되어 있는 244,824개의 MathML 수식을 수집하여 테스트 컬렉션으로 사용하였다. 그리고 제안 시스템의 평가를 두 가지 형태로 진행하였다. 먼저 색인되어 있는 수식 중 임의로 10,000개를 골라 질의식을 구성한 후, 해당 수식이 몇 위에 검색되는지 평가하는 비공개 테스트(closed test)를 진행하였다. 다음으로 색인되어 있지 않은 50개의 질의식을 이용한 공개 테스트(open test)를 진행하였다. 공개 테스트에서 검색 수식의 적합성 평가는 3명의 연구원에 의해 수행되었다. 성능 평가는 식 (4)와 같은 상위 n순위에 대한 정확률(이하 P@N)을 이용하였다.

$$P@N = \frac{n\text{순위내적합수식의수}}{n} \quad (4)$$

4.1 비공개 테스트

비공개 테스트에서 Youssef의 수식 순위화 기법과 각 색인단위에 동일한 가중치를 주는 기법을 이용하여 성능 비교를 하였다. Youssef 수식 순위화 기법은 수식의 연산자, 변수, 함수명에 서로 다른 가중치를 부여하며 수식의 구조는 색인에 고려하지 않는다. 비공개 테스트의 결과는 표 5와 같다. 표 5에서 보는 것처럼 색인된 10,000개의 수식을 질의식으로 하였을 때 동일 가중치 부여 기법과 색인어 별 가중치 기법의 P@1 평균 정확률은 53% 였으며 Youssef의 수식 순위화 기법은 50% 였다. 이와 같은 결과로 수식의 구조가 수식 검색 시스템에서 중요한 색인 단위가 된다는 것을 확인할 수 있다. 비공개 테스트를 분석한 결과 동일 순위를 갖는 수식들이 다수 출현하였다. 표 5의 비공개 테스트에서 검색된 수식이 같은 점수를 갖는 경우에는 검색된 순서에 따라 순위를 부여하였다. 또한 동일 가중치 부여 방법과 제안 방법이 같은 성능을 보였는데, 이는 질의 수식과 같은 수식이 하나만 존재하며 이러한 이유에서 가중치

표 5 비공개 테스트의 평균 P@1

질의식 수	P@1		
10,000	Youssef 순위화	동일 가중치 부여	제안 방법
	0.50	0.53	0.53

표 6 각 색인 단위 별 비공개 테스트의 평균 P@1

	변수만 색인	연산자만 색인	구조만 색인
P@1	0.42	0.37	0.28

보다 검색되는 색인 단위 자체가 중요하기 때문이다.

수학식 색인에 사용된 연산자, 변수, 구조의 중요성을 측정하기 위해 추가 실험을 진행하였다. 표 6은 연산자만 색인한 방법, 변수만 색인한 방법, 구조만 색인한 방법의 비공개 테스트 결과이다. 표에서 보면 구조만 색인한 경우 P@1 값이 가장 낮은 것을 볼 수 있다. 이러한 결과는 표 2에서 구조의 수가 다양하지 않기 때문인 것으로 판단된다. 앞선 실험과 이 실험을 통하여 수식 구조가 갖는 정보량은 높지만 정보를 표현하는데 필요한 수식 구조의 종류가 다양하지 않기 때문에 수식 구조는 변수, 연산자와 함께 색인되어야 한다는 것을 알 수 있다.

추가적으로 색인되어 있는 50개의 수식에서 변수만을 변경한 동일 의미의 새로운 수식을 이용하여 P@N을 측정하는 실험을 진행하였다. 표 7은 색인된 수식의 변수를 변경하여 같은 의미의 수식을 생성한 예이다.

표 8은 각 Youssef 방법과 색인어에 동일 가중치를 부여하는 방법, 본 논문에서 제안한 색인어 별 가중치를 부여하는 방법을 비교한 것이다.

표 8에서 보는 것과 같이 Youssef 순위화 방법, 동일 가중치 부여 방법보다 제안 방법이 가장 높은 성능을 보여주었다. Youssef 방법은 수식의 구조를 색인에 반

표 8 변형된 질의식의 평균 검색 순위

	Youssef 순위화	동일 가중치 부여 방법	제안 방법
평균순위	5,150위	35위	29위

표 7 의미가 같은 수식 예제

색인 수식	변형 수식
$G(\lvert x-y\rvert,z)=\frac{1}{\sqrt{4\pi z}}e^{-\frac{\lvert x-y\rvert^2}{4z}}$	$G(\lvert a-b\rvert,c)=\frac{1}{\sqrt{4\pi c}}e^{-\frac{\lvert a-b\rvert^2}{4c}}$
$(\mu_1-\mu_2^*)^{-n(n-1)}\frac{1}{\Delta(\hat{q_1})\Delta(\hat{q_2})}e^{\frac{i}{2}N(\mu_1+\mu_2^*)Tr(\hat{q_1}+\hat{q_2})}$	$(\alpha_1-\alpha_2^*)^{-m(m-1)}\frac{1}{\Delta(\hat{p_1})\Delta(\hat{p_2})}e^{\frac{i}{2}M(\alpha_1+\alpha_2^*)Tr(\hat{p_1}+\hat{p_2})}$
$S\mapsto\begin{pmatrix}1_n & 0\\ 0 & A\end{pmatrix}S\begin{pmatrix}1_n & 0\\ 0 & B\end{pmatrix}$	$S\mapsto\begin{pmatrix}1_m & 0\\ 0 & \alpha\end{pmatrix}S\begin{pmatrix}1_m & 0\\ 0 & \beta\end{pmatrix}$
(n^2-16)	(A^2-16)
$w=\sum_a D^a T_a,$	$w=\sum_x D^x T_x,$

영한 방법보다 현저히 낮은 성능을 보여주었다. 이는 앞에서 언급한 것과 같이 수식의 구조가 수식 검색 시스템에서 중요한 색인 단위가 된다는 것을 알 수 있다. 또한 동일한 가중치를 부여하는 방법과 제안 방법인 색인어별 가중치 부여 방법을 비교한 결과 구조가 단순한 수식에서는 큰 차이가 없었으나 구조가 복잡한 수식일수록 제안 방법의 성능이 더 높아지는 것을 확인할 수 있었다.

4.2 공개 테스트

공개 테스트를 위해서 정보 검색 분야 대학원생 1명, 검색을 자주 이용하는 학부생 1명, 그리고 검색을 거의 하지 않는 학부생 1명이 실험에 참여하였다. 공개 테스트는 검색 된 상위 N개의 수식 중 질의수식과 유사하다고 판단되는 수식을 정답으로 하였다. 그림 3은 색인되어 있지 않은 50개의 질의식을 이용하여 제안 시스템의 P@N을 측정한 결과를 보여준다.

일반적으로 P@N은 n이 증가할수록 더 높은 정확률을 보인다. 그러나 본 논문에서 사용한 테스트 컬렉션은 비슷한 의미의 수식들을 거의 포함하고 있지 않아서 n이 증가해도 적합식의 수는 거의 증가하지 않았다. 그러므로 n이 증가할수록 P@N이 점차 떨어지는 경향을 보였다. 그림 3에서 보는 것처럼 제안 방법이 동일 가중치 부여 방법보다 평균 6% 정도 높은 정확률을 보였다. 표 9는 공개 테스트에 사용된 수식의 예이다.

표 10은 공개 테스트에 사용된 수식과 검색된 수식의

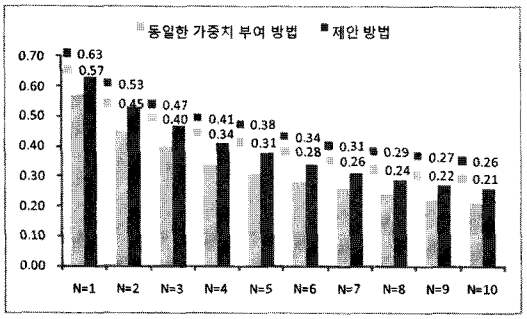

그림 3 제안 시스템의 P@N

표 9 공개 테스트에 사용된 수식 예제

1	$\bar{D}^{ia}\tilde{Q}^{j}$	6	$\rho(z) \in L^1(\mathbb{R}, dz)$
2	$\delta_{\Delta_k=\Delta_{k-1}\delta}, k = 1, \ldots, n,$	7	$\hat{f} \in A \otimes A \otimes A$
3	$\eta \in (0,1)$	8	$F\left(g^{kh}\right)_{jj}^{m} = 0$
4	$S \mapsto \begin{pmatrix} 1_\eta & 0 \\ 0 & A \end{pmatrix} S \begin{pmatrix} 1_\eta & 0 \\ 0 & B \end{pmatrix}$	9	$P_q T_q Q$
5	$gl(n+1 \mid m)$	10	$\lambda = \sum_{i=1}^{r} \lambda_i \Lambda^i = \sum_{i=1}^{r} \lambda^i a_i^{\vee},$

예이다. 질의 식 (1)은 변수가 다수 포함된 수식이다.

이 때 동일 가중치 부여 방법의 검색 결과에서는 질의 수식과 같은 변수들이 다수 포함된 문서들이 상위에 검색되는 것을 확인할 수 있었다. 이는 질의 수식과 의미가 다르기 때문에 유사한 수식으로 판단되지 않았다. 그러나 제안 방법에서는 실제 수식과 검색된 수식의 변수는 다르지만 매우 유사한 구조의 수식이 상위에 검색되는 것을 확인할 수 있었다. 이는 질의 수식과 유사한 수식으로 판단되었으며 실험에서 동일 가중치 부여 방법보다 제안 방법이 더 높은 성능을 보였다. 질의 수식 2, 4는 복잡한 구조의 질의 수식으로 구조에 더 높은 가중치를 주는 제안 방법이 동일 가중치 부여 방법보다 높은 성능을 보였다. 질의 수식 (3)에서는 동일 가중치 부여 방법에서 구조의 가중치가 제안 방법에서 구조의 가중치보다 더 낮기 때문에 오히려 구조가 크게 반영되지 못하였다. 이러한 이유에서 질의 수식의 구조보다 조금 더 복잡한 수식이 동일 가중치 부여 방법의 검색 결과로 나타났다.

5. 결론 및 향후연구

본 논문에서는 수식을 분석하여 연산자, 변수, 구조를 별도로 색인하고 개별 검색을 수행한 후 가중치 합을 통해서 수식 검색의 성능을 향상시키는 방법을 제안하였다. 실험결과에 따르면 질의식과 동일한 색인식이 존재하는 경우에 53%의 P@1을 보였다. 또한 변수만을 변경한 동일 의미의 수식 검색 결과에서도 의미있는 결과를 확인할 수 있었다. 색인에 참여하지 않은 수식을 검색식으로 사용한 실험에서는 63%의 P@1을 보였다.

표 10 공개 테스트의 질의 수식 및 검색된 수식 예제

	질의 수식	제안 방법 검색 예	동일 가중치 부여 방법 검색 예
1	$\bar{D}^{ia}\tilde{Q}^{j}$	$\bar{D}^{ia}\bar{R}^{b}$	$\bar{D}^{ia}\tilde{Q}^{j} = -\bar{D}^{ja}\tilde{Q}^{i}$
2	$\delta_{\Delta_k=\Delta_{k-1}\delta}, k = 1, \ldots, n,$	$d_{\Delta_k=\Delta_{k+1}d}, k = 0, \ldots, n,$	$a_{sj}, s = 1, \ldots, lg, j = 1, \ldots, l-1$
3	$\eta \in (0,1)$	$\tau \in [0,1)$	$\eta \in \left[0, \eta_1\right]$
4	$S \mapsto \begin{pmatrix} 1_\eta & 0 \\ 0 & A \end{pmatrix} S \begin{pmatrix} 1_\eta & 0 \\ 0 & B \end{pmatrix}$	$M_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, M_2 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, M_3 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, M_4 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	$S \mapsto ASB$

아직 해결하지 못한 향후 연구 과제는 다음과 같다. 질의식의 형태를 파악하여 변수의 수가 임계치보다 큰 경우에는 구조와 연산자에 더 높은 가중치를 두어서 수식의 의미가 변질되지 않는 방법에 대해서 연구할 계획이다. 또한 수식 주위의 텍스트를 분석하여 수식과 텍스트 사이의 상관 관계를 파악하고, 이를 기반으로 키워드를 이용하여 수식을 검색하는 방법을 연구할 예정이다.

참 고 문 헌

[1] Mathematical Markup Language, http://www.w3.org/math

[2] M. Adeel, H. S. Cheung and S. H. Khiyal, "MATH GO! Prototype of a Content Based Mathematical Formula Search Engine," *Journal of Theoretical and Applied Information Technology*, vol.4, no.10, pp.1002-1012, 2008.

[3] J. Misutka, L. Galambos, "Extending Full Text Search Engine for Mathematical Content," *Proceedings of Towards Digital Mathematics Library*, pp.55-67, 2008.

[4] A. S. Youssef, "Relevance Ranking and Hit Description in Math Search," *Mathematics in Computer Science*, vol.2, no.2, pp.333-353, 2008.

[5] J. M. Ponte, W. B. Croft, "A Language Modeling Approach to Information Retrieval," *Proceedings of ACM SIGIR*, pp.275-281, 1998.

[6] http://arxmliv.kwarc.info/files/math-ph/papers/

[7] D. Hiemstra, "Using Language Models for Information Retrieval," Ph.D. Thesis, *Centre for Telematics and Information Technology, University of Twente*, ISBN 90-75296-05-3, 2001.

[8] J. S. Shin, S. H. Lee, H. S. Kim, "Mathematical Equation Retrieval Based on Properties of Mathematical Symbols," *Proceedings of the 36th KIISE Fall Conference*, vol.36, no.2(C), pp.188-193, 2009. (in Korean)

[9] M. E. Altamimi, A S. Youssef, "A More Canonical Form of Content MathML to Facilitate Math Search," *Proceedings of the 2007 Extreme Markup Languages Conference*, 2007.

신 준 수

2009년 강원대학교 컴퓨터정보통신공학부 학사. 2009년~현재 강원대학교 컴퓨터정보통신공학전공 석사과정. 관심분야는 정보검색 시스템, 감정분류 시스템, 대화 인터페이스 시스템

김 학 수

1996년 건국대학교 전자계산학과 학사. 1998년 서강대학교 컴퓨터학과 석사. 2003년 서강대학교 컴퓨터학과 박사. 2004년 University of Massachusets, Amherst 박사후연구원. 2005년 한국전자통신연구원 선임연구원. 2006년~현재 강원대학교 컴퓨터정보통신공학전공 조교수. 관심분야는 한국어정보처리, 생략 및 대용어 처리, 대화 인터페이스 시스템, 정보검색 시스템, 질의응답 시스템