# 퍼지 함수에 의한 질의어 확장과 문서 분류 알고리즘

## (An Algorithm of Documents Classification and Query Extension using Fuzzy Function)

은 희 주 † 하 얀 †† 김 용 성 †††
(Hye-Ju Eun) (Yan Ha) (Yong-Sung Kim)

**요 약** 웹 기반 검색 시스템에서 사용자의 관심이 많은 문서를 선별하여 제공하기 위해 프로파일이나 시소러스에 관한 연구가 이루어지고 있다. 그러나, 프로파일이나 시소러스를 구축하고 유지보수 하는데 많은 시간과 노력이 필요하며, 특히 구축된 시소러스에 대해 구조화 및 적합성의 문제가 있다. 따라서, 이러한 문제점을 극복하고자 본 논문에서는 문서에서 추출한 용어 빈도를 문서에서 용어의 중요 정도로 사상시키기 위해 시그모이드 멤버 쉽 함수를 적용한다. 또한, 이 중요 정도에 따라 질의어를 확장하고 의미적으로 연결된 문서를 동일한 문서 집단으로 분류할 수 있는 알고리즘을 제안하여 사용자의 선호도가 반영된 문서를 선별하고 제공하고자 한다.

***Abstract*** Profiles and thesaurus has been studied to select and provide for disseminating documents in which the user is interested based on the web information system. But these methods need much time and overhead to be built accurately and building a thesaurus especially presents problems of construction and fitness. To overcome these problems, in this paper, the sigmoid membership functions applied to transform occurrence-frequency of keywords extracted in documents into degrees of importance. Also, we propose an algorithm to extend queries and classify into the same cluster connected documents of similar meaning according to degree of importance. As result, we are able to select and to provide user-interested documents.

## 1. 서 론

현재 네트워크의 기술이 급격히 발달하고, 인터넷이 전 세계에 걸쳐 보급됨에 따라 사용자는 산재된 많은 양의 정보를 손쉽게 얻을 수 있으나 실제 이런 정보들 중에서 사용자의 관심도가 많이 반영된 문서를 선별하여 제공하는 검색 시스템을 구현하는데는 어려움이 있다. 이를테면 , 질의어가 문서 내에서 발생 여부에 따라 검색 결과를 제공하는 키워드 매칭 시스템의 경우, 동의어 등의 처리가 제대로 되지 않아 무분별한 문서 제공이 이루어지기 때문에 사용자는 많은 시간과 노력을 투자하여 여러 번 피드백 과정을 거쳐야 하는 번거로움이 있다. 이러한 단점을 해결하고자 프로파일(profile)이나 시소러스(thesaurus)를 이용한 질의어 확장 방법에 관한 연구들이 활발히 진행되고 있으나, 이런 경우 프로파일이나 시소러스 구축에 많은 시간이 필요할 뿐만 아니라 구축된 시소러스에 대해 구조화 및 적합성의 문제가 발생한다[1]. 구체적으로 기술하면, 용어간 관계성 불일치 문제가 발생할 경우 최소의 논리합 정규 변환으로 인해 검색 시간이 증가될 뿐만 아니라 거리 알고리즘에서 NOT 연산자 사용에 따라 비효율적인 문제가 발생하게 된다. 또한, Min과 Max함수를 이용한 OR 연산의 비효율성에 관한 문제도 발생한다[2].

따라서, 시소러스에 대한 이러한 문제점을 해결하고 개념 정보들이 속해 있는 용어들의 유사도(similarity) 및 개념 거리(concept distance)를 이용하여 개념 정보

† 학생회원 : 전북대학교 컴퓨터통계정보학과
hjeun@cs.chonbuk.ac.kr
†† 정 회 원 : 경인여자대학 멀티미디어정보전산학부 교수
yanha@object.cse.cau.ac.kr
††† 종신회원 : 전북대학교 컴퓨터통계정보학과 교수
yskim@moak.chonbuk.ac.kr
논문접수 : 2000년 2월 21일
심사완료 : 2000년 11월 17일

(concept information)를 구성하는 연구의 일환으로, 본 연구에서는 사용자 관심을 표현하는 질의어와 문서에서 추출한 색인어간의 유사 정도를 [0, 1]사이의 퍼지 값으로 사상시켜 질의어를 확장하고자 한다.

또한, 동일한 문서뿐만 아니라 다른 문서에서 어떤 단어나 구가 동시에 발생하는 빈도가 높을수록 그 단어를 포함한 문서끼리는 서로 의미적으로 연결된다는 사실에 기반을 두고 질의어와 직접 매칭 되지 않은 문서까지 검색할 수 있도록 퍼지 논리(fuzzy logic)를 기반으로 문서를 분류하는 알고리즘을 제안한다.

본 논문의 구성은 2장에서 관련 연구에 대해 살펴보고, 3장에서는 퍼지 집합 및 관계를, 4장에서는 본 연구에서 제안하는 퍼지 논리를 기반으로 한 질의어 확장과 문서 분류 알고리즘을 제안하고, 5장에서는 제안된 알고리즘을 평가·검증하기 위해서 실험과 평가를 한다. 끝으로, 6장에서는 결론 및 향후 연구 과제에 대해서 논의한다.

## 2. 관련 연구

최근에 많이 연구되고 있는 질의어 확장에 대한 연구는 개인, 그룹의 프로파일을 이용하는 기법과 사용자 피드백에 의한 기법이 있다. 프로파일을 작성하여 사용하는 경우는 사용자가 질의를 입력하면 이미 사용자의 관심 분야에 관련된 키워드들로 작성된 프로파일이나 시소러스를 참조하여 질의어를 확장하게 된다[3].

프로파일은 사용자 식별자(user id)와 관심분야를 나타내기 위해 사용하는 키워드들의 용어 벡터(term vector)와 선호도(preference)를 반영하기 위한 키워드의 가중치 벡터(weighted vector)로 구성되는데, 사용자의 관심도가 변경될 때에는 프로파일이 갱신 될 수 있도록 제안되고 있다.

사용자 피드백에 의한 질의어 확장 기법은 사용자가 탐색된 결과를 보고 질문을 수정하고 반복적인 탐색을 통해 관련성 높은 문서를 검색하는 기법으로 이 경우 사용자는 시스템의 도움을 받아 데이타베이스를 탐색하기 전에 보다 좋은 질의어를 입력하거나 질의어를 재생성한다[4]. 여기서, 시스템이 제공하는 정보는 탐색 대상의 문서집합에서 시소러스부터 원래의 질의어에 대한 동의어뿐만 아니라 기타 관련 용어 리스트, 질의어가 데이타베이스 내에서 출현한 빈도 그리고 이미 검색된 적합한 문서에서 추출한 색인어에 관한 정보 등이다.

위와 같은 기법들을 이용하면 사용자는 검색하고자 하는 문서에 대해 질의어를 보다 정확하게 표현할 수 있고 키워드 매칭(keyword matching)에 의한 문서뿐만 아니라 관련성 있는 문서까지 검색 할 수 있는 장점이 있으나 프로파일이나 시소러스 구축에 있어서 많은 시간과 노력이 필요하며 동적으로 이들을 유지보수하기가 무척 어려운 단점이 있다.

또한, 질의어 확장과 더불어 정보검색 분야에서 활발히 연구되고 있는 내용은 문서를 분류(classification)하는 작업이다. 문서분류는 정해진 분류체계 하에서 분류하고자 하는 각 문헌들을 가장 적합한 카테고리에 배정함으로써 문서를 집단화(categorization)하는 작업이다. 그러나 현재까지 문서분류 기법들은 검색이 시작되기 전에 클러스터(cluster)가 형성되어야 하는 것과 클러스터의 내용과 센트로이드(centroid)가 계속 변하는 단점을 가지고 있어 많은 양의 문서에 적용하기가 어렵다[5]. 특히, 벡터 유사도를 이용한 문서 분류 방법은 키워드간의 동의어와 불용어 처리가 어렵고 사용자의 관심을 표현하기 위한 방법이 역 문헌빈도(invert document frequence)에 의한 키워드 추출에 한정되어 있을 뿐만 아니라, 학습과정이 비교적 단순하여 사용자의 관심도를 충분히 반영하지 못한다는 문제점이 있다.

따라서, 본 연구에서는 이들 제한점을 극복하고자 퍼지 논리를 기반으로 용어의 사용 빈도에 의한 질의어 확장 및 문서 분류 알고리즘을 제안하고자 한다.

## 3. 퍼지 함수와 퍼지 관계

본 장에서는 퍼지 집합을 나타내기 위한 소속 함수(membership function)와 퍼지 관계성을 살펴본다.

### 3.1 소속 함수

퍼지 집합 $A$가 임의의 $X=\{x\}$에 대하여 [0,1] 값으로 표현되기 위해서는 $x=x_0$에 대해 집합 $A$의 소속 정도(membership degree)를 나타내는 소속 함수, $\mu_A: X \rightarrow [0,1]$으로 정의된다.

이진 퍼지 관계(binary fuzzy relation)는 전체집합 $A$의 임의의 퍼지 집합 $X$, $Y$에 대해 퍼지 집합 $X$의 임의의 원소 $x$와 퍼지 집합 $Y$의 임의의 원소 $y$ 사이에 관계를 순서쌍 $(x, y)$으로 나타내고 $(x, y)$의 모임을 관계 $R$로 표현한다. 다음 표 1는 이진 퍼지 관계를 구성하는 퍼지 집합 $X$와 $Y$의 원소로 이루어진 퍼지 행렬을 나

표 1 이진 퍼지 행렬

| Y \ X | $y_1$ | $y_2$ | $y_3$ | ⋯ |
|---|---|---|---|---|
| $x_1$ | $\mu(x_1, y_1)$ | $\mu(x_1, y_2)$ | $\mu(x_1, y_3)$ | ⋯ |
| $x_2$ | $\mu(x_2, y_1)$ | $\mu(x_2, y_2)$ | $\mu(x_2, y_3)$ | ⋯ |
| ⋯ | ⋯ | ⋯ | ⋯ | ⋯ |

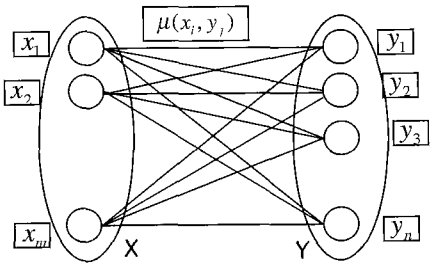


그림 1 이진 퍼지관계 그래프

타낸다.

표 1을 그래프로 표현하면 그림 1과 같다.

그림 1에서 관계 $R$를 정의하면, $X=\{x\}$, $Y=\{y\}$에 대해, $X\times Y=\{(x,y)\}$의 소속 함수 $\mu_R$: $X\times Y\rightarrow[0,1]$로 정의할 수 있다.

이런 퍼지 값으로 사상시키는 소속 함수로 가장 많이 사용하는 대표적인 예는 $S$-소속 함수와 시그모이드 함수가 있다.

3.1.1 S-소속 함수

S-소속 함수는 $[x_m, x_M]$에서 임의의 퍼지 집합 $X$를 표현하고자 할 때 사인함수를 이용하여 소속함수를 생성할 수 있다. 만약, $0\leq\theta\leq\frac{\pi}{2}$에서 사상 함수 $S_1=\sin\theta$이고, $x_m\leq x\leq x_M$에서 $\theta=\frac{\pi}{2}(\frac{x-x_m}{x_M-x_n})$으로 주어지면 소속 함수 $S_1$는 $\mu_{S_1}=S_1(x,x_m,x_M)=\sin\frac{\pi}{2}(\frac{x-x_m}{x_M-x_m})$이다.

따라서, 그림 2와 같이 사인함수를 이용한 소속 함수는 퍼지 값이 0.5인 점을 기준으로 정확히 대칭인 함수이며 규칙적인 값을 갖도록 유도된다[6].

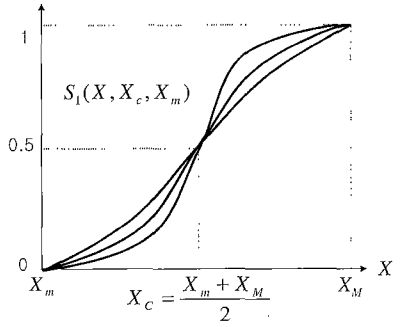


그림 2 S-소속 함수

3.1.2 시그모이드 소속 함수

본 연구에서 퍼지 소속함수로 정의한 시그모이드 (sigmoid) 함수는 다음의 세 가지 특징을 만족한다. 즉, 입력 값에 무관하게 항상 [0, 1]사이의 퍼지 값을 갖고 S자 형태의 단조 증가 형태를 이루고 또한, 임계값(critical value)을 갖는 퍼지 소속 함수이다[7].

1) $\sigma: R^+\rightarrow[0,1]$
2) $\sigma(F_1)>\sigma(F_2)\Leftrightarrow F_1>F_2$
3) $\frac{d^2(\sigma)}{dF^2}\geq 0\Leftrightarrow F\leq T_F$ and $\frac{d^2(\sigma)}{dF^2}\leq 0\Leftrightarrow F\geq T_F$

위의 조건들을 만족하는 시그모이드 함수에 대한 그래프는 그림 3과 같이 나타낼 수 있다.

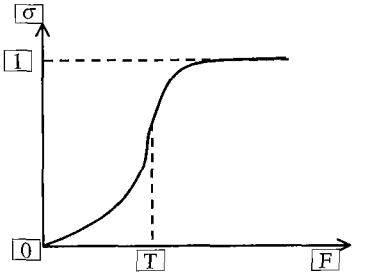


그림 3 시그모이드 함수

본 논문에서는 각 문서에서 추출한 색인어의 발생 영역에 따른 빈도를 문서의 내용과 의미를 나타내는 중요 정도로 사상시키기 위해서 그림 4와 같이 각기 다른 임계값을 갖는 시그모이드 함수들을 정의하고, 또한 이러한 임계값은 질의어 확장과 문서 분류에 적용된 $\alpha$-cut의 알파($\alpha$) 값에 이용된다.

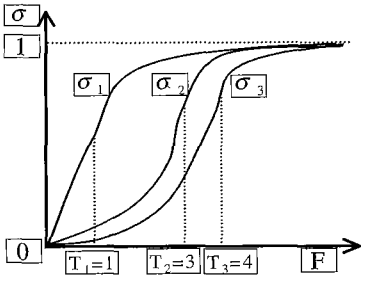


그림 4 시그모이드 함수 정의

### 3.2 $\alpha$-cut

$\alpha$-cut은 소속 함수의 [0,1]사이의 값에서 임의의 $\alpha$ ($0\leq\alpha\leq1$)값이 되는 함수 값에 대한 퍼지 상태 변수의 구간을 나타낸다. 이 $\alpha$-cut은 퍼지 집합의 원소들에 대해 집합에 속할 기준을 정의할 때 사용된다. 임의의 $X$을 원소로 하는 퍼지 집합 $A$에 대해서 임의의 $\alpha\in[0,$

1] 값을 가진 $\alpha$-cut을 적용한 퍼지 집합 $A_\alpha$는 다음과 같이 정의한다[7].

$A_\alpha = \{x | A(x) \geq \alpha\}$

따라서, 퍼지 집합 $A_\alpha$는 퍼지 집합에 속할 소속정도의 값이 $\alpha$값 이상으로 이루어진 집합이다.

다음 그림 5는 사다리꼴 소속 함수에 대한 그래프로 퍼지 집합을 표시하며 2개의 $\alpha$-cut을 적용했을 때 생성되는 퍼지 집합간의 포함 관계를 나타낸다.

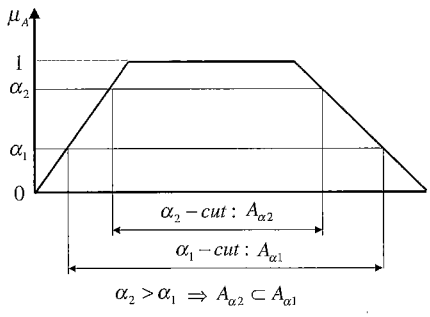


그림 5 $\alpha$-cut

### 3.3 유사관계

임의의 퍼지 집합 $X$에서 각각의 원소 $x$, y에 대하여 동치관계를 만족하는 집합 $X$의 모든 원소들을 포함하는 집합 $A_x$는 다음과 같이 관계로 표현할 수 있다.

$$A_x = \{y \mid \langle x, y \rangle \in R(X, X)\}$$

집합 $X$가 퍼지 집합일 때, 임의의 $x, y, z \in X$에 대하여 정의된 퍼지 관계 $R \subseteq X \times X$이 다음과 같이 반사관계, 대칭관계, 전이관계가 정의될 때 유사관계($\cong$)(similarity relation)가 성립된다[8].

반사관계 : $\mu_{\cong}(x, x) = 1$
대칭관계 : $\mu_{\cong}(x, y) = \mu_{\cong}(y, x)$
전이관계 : $\mu_{\cong}(x, z) \geq \min\{\mu_{\cong}(x, y), \mu_{\cong}(y, z)\}$
단, $\mu$ :소속함수

퍼지 집합에서 유사관계를 만족하는 임의의 원소들은 퍼지 집합에 대한 소속정도의 값을 부여받고 특정한 소속정도의 값에 따라 유사한 원소들로 이루어진 유사클래스(similarity class)를 생성하여 그룹화 할 수 있다.

### 3.4 호환 관계

퍼지 관계 중 호환관계($\approx$)(tolerance, compatibility relation)는 다음과 같이 반사와 대칭 성질을 만족하고 전이관계는 만족하지 않는다.

반사관계 : $\mu_{\approx}(x, x) = 1$
대칭관계 : $\mu_{\approx}(x, y) = \mu_{\approx}(y, x)$

퍼지 집합에서는 퍼지 관계 $R$이 퍼지 호환 관계를 만족할 때 특정한 소속정도 $\alpha$값을 선택, $\alpha$-cut을 적용하여 호환 클래스 집합 $A_\alpha$를 생성한다.

$\alpha$-호환 클래스는 임의의 $x, y$에 대하여 이들 관계가 $\alpha$값 이상이면 $X$의 부분집합으로 구성된다. 이렇게 구성된 모든 호환 클래스들을 최대 호환 클래스 또는 완전 $\alpha$-cover라고 한다[9].

## 4. 질의어 확장과 문서 분류

본 장에서는 퍼지 시그모이드 소속 함수를 기반으로 한 질의어 확장과 문서 분류 알고리즘을 제안한다.

### 4.1 시그모이드 함수를 이용한 퍼지 소속함수

키워드는 문서의 내용을 대표할 수 있는 중요한 단어나 단어 구로 구성되며 임의의 문서간에 동일한 키워드의 출현빈도가 높을수록 문서들 사이는 서로 내용적으로 연결되어 있다고 볼 수 있다.

따라서, 본 논문에서는 특정 분야별 문서 집합에 대하여 각 문서에서 추출한 키워드의 발생 빈도를 문서에 대한 중요 정도로 사상시키기 위해 소속함수를 이용하고 전체 단어에 대한 키워드의 발생 확률(probability)값이 아닌 키워드가 전체 문서의 내용을 대표할 가능(possibility) 정도의 값으로 나타낸다.

본 논문에서 키워드들 사이의 퍼지 관계성을 정의하기 위해서 50개의 논문으로 구성된 실험 집단에서 추출한 150개 키워드를 추출하여 실험하였고, 그 결과를 이용해서 다음과 같이 3개의 소속 함수를 정의하고, 또한 150개의 키워드 중 임의의 10개의 키워드에 대한 중요 정도를 이용하여 질의어 확장과 문서분류 과정을 보여준다.

첫째, 문서에서 추출한 키워드(단어나 단어 구)가 문서의 타이틀이나 키워드 집합(keyword set)에서 발생되었을 때 키워드 발생 빈도에 대한 문서에서의 중요 정도는 그림 4의 시그모이드 함수($\sigma_1$)에 의해 표 2와 같이 구할 수 있다.

표 2 타이틀, 키워드 집합에서의 소속 정도

| F | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| $\sigma_1$ | 0 | 0.6 | 0.9 | 0.99 | 1 |

위의 표 2를 시그모이드 소속 함수를 적용하면 그림 6와 같다.

둘째, 키워드가 문서의 요약과 결론 부분에 발생되었을 경우 빈도에 대한 소속 정도는 그림 4의 시그모이드 함수($\sigma_2$)에 의해 표 3과 같이 구할 수 있다.

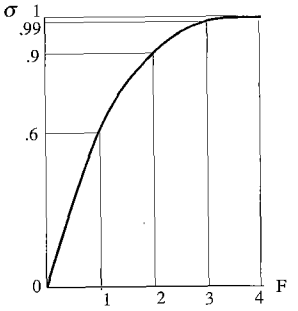


그림 6 소속함수 ( $S_1$)

표 3 요약과 결론에서의 소속 정도

| F | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| $\sigma_2$ | 0 | 0.1 | 0.25 | 0.7 | 0.92 | 0.97 | 1 |

위 표를 시그모이드 함수로 나타내면 그림 7와 같다.

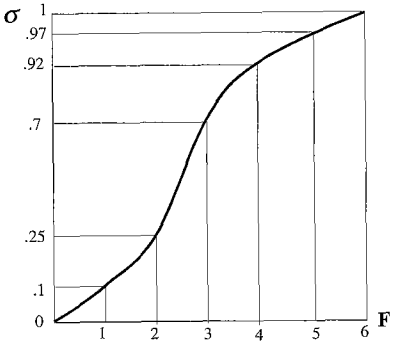


그림 7 소속함수( $S_2$)

셋째, 문서의 본문에서 발생되었을 때 그림 4의 시그모이드함수( $\sigma_3$)에 의해 표 4와 같이 구할 수 있다.

표 4 본문에서의 소속 정도

| F | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sigma_4$ | 0 | 0.09 | 0.25 | 0.4 | 0.7 | 0.85 | 0.9 | 0.95 | 0.98 | 0.99 | 1 |

그림 8을 살펴보면 발생 빈도가 4일 때 임계값을 가지게 되어 소속 정도의 값이 급격하게 증가하고 있음을 보여주고 있다.

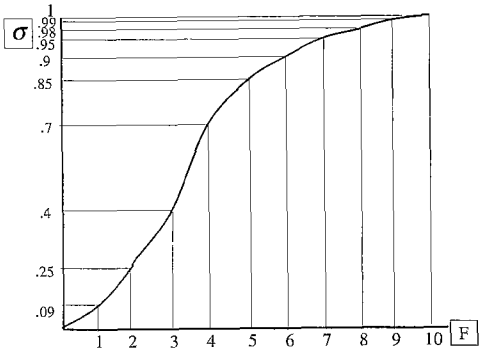


그림 8 소속함수( $S_3$)

각각의 문서에서의 키워드 발생 영역별 빈도에 대한 소속 정도를 이용하여 키워드가 임의의 한 문서에서의 중요 정도는 (식 1)을 적용하여 구할 수 있다.

$$\mu_{ij}(x) = \max\{\min\{\mu_T(x), \mu_A(x)\}, \min\{\mu_A(x), \mu_S(x)\}, \min\{\mu_S(x), \mu_T(x)\}\} \quad \text{---(식 1)}$$

$\mu_{ij}(x)$: 문서 $j$에서 키워드 $i$의 중요 정도
$\mu_T(x)$: 문서 $j$에 대한 타이틀과 키워드 집합의 키워드 $i$의 중요 정도
$\mu_A(x)$: 문서 $j$에 대한 요약이나 결론 영역의 키워드 $i$의 중요 정도
$\mu_S(x)$: 문서 $j$에 대한 본문의 키워드 $i$의 중요 정도

### 4.2 유사 클래스를 이용한 질의어 확장

문서에서 추출한 키워드의 발생 위치에 따라 문서의 내용을 의미적으로 대표하는 정도가 다르므로 문서에 대한 키워드의 중요도를 서로 다른 소속함수를 정의하여 키워드의 가중치를 고려하여 퍼지 유사 관계를 만족하는 키워드들 사이의 관련 정도에 따라 유사 클래스를 생성한다. 생성된 유사 클래스는 사용자의 질의어와 유사한 키워드들로 구성되어 질의어를 확장시켜 내용 기반의 검색을 하도록 한다. 즉, 기존의 부울 검색 시스템이 제공하는 키워드 직접 매칭에 의한 검색 방법의 단점인 자연어의 동의어 관계를 파악하여 불필요한 문서를 필터링(filtering)한다.

다음은 특정 분야에 관련된 문서 집합 $D=\{D_1, D_2, D_3, ..., D_n\}$에서 추출한 키워드 집합 $K=\{k_1, k_2, k_3, ..., k_m\}$에 대하여, 각 문서에 대한 각 키워드들의 발생 영역별 빈도를 이용한 질의어 확장 과정을 단계별로 제시한다.

**[Step 1] 문서에서 추출한 키워드의 발생 영역별 출현 빈도를 구한다.**

각 키워드들에 대한 타이틀과 키워드 집합에서의 발생빈도는 표 5와 같다.

표 5 타이틀이나 키워드집합에서 키워드의 빈도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0 |
| 2 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 6 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 7 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 9 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 10 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

또한, 키워드들에 대한 요약과 결론 영역에서의 출현 빈도는 표 6과 같다.

표 6 요약과 결론 부분에서 키워드의 빈도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 1 | 0 | 5 | 0 | 0 | 1 | 0 | 3 | 0 |
| 2 | 0 | 1 | 3 | 0 | 2 | 4 | 0 | 0 | 1 | 2 |
| 3 | 1 | 0 | 0 | 2 | 4 | 0 | 3 | 1 | 0 | 1 |
| 4 | 1 | 0 | 1 | 4 | 0 | 0 | 0 | 4 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 1 |
| 6 | 3 | 1 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 2 |
| 7 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 1 | 0 | 0 |
| 8 | 0 | 0 | 1 | 0 | 5 | 0 | 5 | 0 | 1 | 0 |
| 9 | 1 | 4 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 1 |
| 10 | 0 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |

마지막으로 각 키워드들에 대한 본문 영역에서의 출현 빈도는 다음 표 7과 같다.

표 7 본문에서 키워드의 빈도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 7 | 3 | 2 | 7 | 1 | 0 | 3 | 0 | 6 | 2 |
| 2 | 1 | 2 | 8 | 2 | 3 | 6 | 1 | 2 | 3 | 5 |
| 3 | 4 | 1 | 0 | 5 | 8 | 1 | 6 | 2 | 1 | 1 |
| 4 | 3 | 0 | 3 | 8 | 1 | 1 | 2 | 8 | 2 | 2 |
| 5 | 0 | 2 | 0 | 0 | 1 | 0 | 3 | 0 | 6 | 3 |
| 6 | 3 | 4 | 4 | 0 | 6 | 3 | 0 | 0 | 2 | 4 |
| 7 | 1 | 0 | 2 | 3 | 0 | 6 | 0 | 3 | 0 | 0 |
| 8 | 0 | 1 | 3 | 0 | 3 | 0 | 8 | 1 | 3 | 0 |
| 9 | 1 | 7 | 1 | 3 | 0 | 1 | 0 | 5 | 0 | 4 |
| 10 | 0 | 2 | 0 | 2 | 1 | 0 | 0 | 1 | 2 | 0 |

**[Step 2] 각 키워드의 발생 영역별 빈도를 문서에서의 중요 정도의 값으로 사상시킨 시그모이드 소속 함수 값을 계산한다.**

다음 표 8은 타이틀 부분과 키워드 집합에서의 발생한 키워드의 중요 정도를 보여주고 있다.

표 8 타이틀과 키워드 집합에서의 소속 정도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.6 | 0 | 0 | 0.9 | 0 | 0 | 0 | 0 | 0.6 | 0 |
| 2 | 0 | 0 | 0.6 | 0 | 0 | 0.6 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0.6 | 0.6 | 0 | 0 | 0.6 | 0 | 0 |
| 4 | 0.6 | 0 | 0 | 0 | 0 | 0 | 0 | 0.9 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.6 | 0.6 |
| 6 | 0.6 | 0.6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.9 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0.6 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0.6 | 0 | 0 | 0 | 0 | 0 | 0.6 | 0 |
| 9 | 0 | 0.6 | 0 | 0.6 | 0 | 0 | 0 | 0.6 | 0 | 0.6 |
| 10 | 0 | 0 | 0 | 0 | 0.6 | 0 | 0 | 0 | 0 | 0 |

또한, 표 9는 요약과 결론에서의 발생한 키워드의 소속 정도를 보여주고 있다.

표 9 요약과 결론 부분에서의 소속 정도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .92 | .1 | 0 | .97 | 0 | 0 | .1 | 0 | .7 | 0 |
| 2 | 0 | .1 | .7 | 0 | .3 | .92 | 0 | 0 | .1 | .3 |
| 3 | .1 | 0 | 0 | .3 | .92 | 0 | .7 | .1 | 0 | .1 |
| 4 | .1 | 0 | .1 | .92 | 0 | 0 | 0 | .92 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0 | .92 | .1 |
| 6 | .7 | .1 | .3 | 0 | .1 | .1 | 0 | 0 | 0 | .3 |
| 7 | 0 | 0 | 0 | .1 | 0 | .7 | 0 | .1 | 0 | 0 |
| 8 | 0 | 0 | .1 | 0 | .97 | 0 | .97 | 0 | .1 | 0 |
| 9 | .1 | .92 | 0 | .1 | 0 | 0 | 0 | .3 | 0 | .1 |
| 10 | 0 | .3 | 0 | 0 | 0.1 | 0 | 0 | 0 | 0 | 0 |

표 10 본문에서의 키워드의 소속 정도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .95 | .4 | .25 | .95 | .09 | 0 | 0.4 | 0 | .9 | .25 |
| 2 | .09 | .25 | .98 | .25 | .4 | .9 | .09 | .25 | .4 | .85 |
| 3 | .7 | .09 | 0 | .85 | .98 | .09 | .9 | .25 | .09 | .09 |
| 4 | .4 | 0 | .4 | .98 | .09 | .09 | .25 | .98 | .25 | .25 |
| 5 | 0 | .25 | 0 | 0 | .09 | 0 | .4 | 0 | .9 | .4 |
| 6 | .4 | .7 | .7 | 0 | .9 | .4 | 0 | 0 | .25 | .7 |
| 7 | .09 | 0 | .25 | .4 | 0 | .9 | 0 | 0.4 | 0 | 0 |
| 8 | 0 | .09 | .4 | 0 | .4 | 0 | .98 | .09 | .4 | 0 |
| 9 | .09 | .95 | .09 | .4 | 0 | .09 | 0 | .85 | 0 | .7 |
| 10 | 0 | .25 | 0 | .25 | .09 | 0 | 0 | .09 | .25 | 0 |

마지막으로 표 10는 본문에서 발생한 키워드의 소속 정도를 보여주고 있다.

**[Step 3] 각 키워드의 문서에서의 평균 소속 정도를 구한다.**

(식 1)를 적용하여 각각의 키워드가 문서의 주제 내용을 얼마나 반영하고 있는지를 계산한 값은 표 11과 같다.

표 11 문서에서의 각 키워드의 소속 정도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .92 | .1 | 0 | .9 | 0 | 0 | .1 | 0 | .7 | 0 |
| 2 | 0 | .1 | .6 | 0 | .3 | .9 | 0 | 0 | .1 | .3 |
| 3 | .1 | 0 | 0 | .6 | .92 | 0 | .7 | .1 | 0 | 0 |
| 4 | .4 | 0 | 0 | .92 | 0 | 0 | 0 | .92 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | .1 | 0 | .6 | .4 |
| 6 | .6 | .6 | .3 | 0 | .1 | .1 | 0 | 0 | 0 | .7 |
| 7 | 0 | 0 | 0 | .1 | 0 | .6 | 0 | .1 | 0 | 0 |
| 8 | 0 | 0 | .4 | 0 | .4 | 0 | 0 | 0 | .4 | 0 |
| 9 | .09 | .92 | 0 | 0 | 0 | 0 | 0 | .6 | 0 | .6 |
| 10 | 0 | .25 | 0 | 0 | .09 | 0 | 0 | 0 | 0 | 0 |

**[Step 4] 키워드들 사이의 유사정도를 이용하여 질의어를 확장한다.**

키워드의 발생영역에 따라 각각 다른 시그모이드 함수를 적용하여 가중치를 고려한 퍼지 값은 문서에서의 키워드의 중요 정도를 나타내고, 또한 이 퍼지 값이 퍼지 유사 관계를 만족할 때 유사정도가 임의의 $\alpha$ 값보다 큰 키워드로 구성되는 유사 클래스를 생성하여 질의를 확장한다.

먼저, 키워드 사이의 유사 정도를 표현하기 위해 두 개의 퍼지 집합 사이의 동치관계를 논리적 동치에 가장 많이 표현되는 불리언 대수를 적용하여 다음 (식 2)와 같이 표현한다.

$$A \equiv B \cong (A \wedge B) \vee (\neg A \wedge \neg B) \quad \text{-------} \ (\text{식 } 2)$$

또한, max-min 연산과 대수의 노름(norms)을 사용하여 (식 3)과 같이 퍼지 소속함수를 사용한 식으로 유도하여 본 논문에서 적용할 수 있도록 키워드들 사이의 유사 정도를 퍼지 값으로 표현한다.

$$\mu_{A \cong B}(x) = \max\{\min\{\mu_A(x), \mu_B(x)\}, \min\{1-\mu_A(x), 1-\mu_B(x)\}\} \text{--(식 3)}$$

$\mu_A(x)$:임의의 원소 $x$가 퍼지 집합 A에 속할 정도

$\mu_B(x)$:임의의 원소 $x$가 퍼지 집합 B에 속할 정도

따라서, 위의 (식 3)에서 볼 수 있듯이 키워드 집합 사이의 유사 관계성은 각 문서집합을 구성하고 있는 키워드들의 소속 정도에 의해 표현될 수 있으며, 문서 집합에서 키워드들 사이의 유사 정도의 값은 (식 4)와 같이 전체 문서에 대한 평균 소속 정도로 표현 할 수 있다.

$$\mu_{ij} = \mu_{w_i \equiv w_j} = \frac{1}{d} \sum_{k=1}^{d} \mu_{w_i \equiv w_j}(D_k) \text{--(식 4)}$$

$\mu_{ij}$ : 키워드 $i, j$ 사이의 유사 정도

$d$ : 전체 문서의 개수

$\mu_{w_i \equiv w_j}(D_k)$ : 문서 $k$에서 키워드 $i, j$ 사이의 유사 정도

(식 4)를 이용하여 키워드들 간의 유사 관계는 표 12와 같다.

표 12 키워드간의 유사 정도

| W \ W | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | .68 | .69 | .56 | .51 | .53 | .61 | .51 | .41 | .65 |
| 2 | .68 | 1 | .67 | .57 | .33 | .45 | .45 | .55 | .71 | .50 |
| 3 | .69 | .57 | 1 | .49 | .50 | .72 | .42 | .53 | .52 | .63 |
| 4 | .56 | .57 | .49 | 1 | .79 | .40 | .40 | .41 | .35 | .68 |
| 5 | .51 | .33 | .50 | .79 | 1 | .38 | .50 | .66 | .57 | .67 |
| 6 | .53 | .45 | .72 | .40 | .38 | 1 | .76 | .55 | .54 | .46 |
| 7 | .61 | .45 | .42 | .40 | .50 | .76 | 1 | .69 | .50 | .48 |
| 8 | .51 | .55 | .53 | .41 | .66 | .55 | .69 | 1 | .48 | .89 |
| 9 | .41 | .71 | .52 | .35 | .57 | .54 | .50 | .48 | 1 | .59 |
| 10 | .46 | .50 | .63 | .68 | .67 | .46 | .48 | .89 | .59 | 1 |

표 12의 키워드들 사이의 유사 정도 값은 퍼지 관계성 중 유사 관계를 만족하므로 $\alpha$-cut을 적용하여 유사 클래스를 생성한다. 그림 9는 0.6-cut을 적용했을 때 생성된 유사 클래스를 나타낸다.

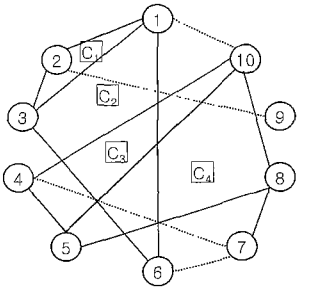


그림 9 0.6-cut 일 때의 유사 클래스

그림 9에서 생성된 유사 클래스는 키워드 집합 $K_1=\{1,2,3\}$, $K_2=\{1,3,6\}$, $K_3=\{4,5,10\}$, $K_4=\{5,8,10\}$와 같은 4개의 집합 사이의 의미적 연결성을 보여준 것으로 사용자의 관심도를 보다 많이 반영되어 결과적으로 질의어 확장을 할 수 있다.

### 4.3 호환 클래스를 이용한 문서분류

일반적으로 임의의 단어나 단어구가 동시 발생 빈도가 많으면 많을수록 문서들은 서로 의미적으로 연결된

다. 본 절에서는 퍼지 관계성을 적용하여 문서간의 의미상 연결 정도를 측정, 문서들을 관련 있는 문서끼리 집단화한다.

문서를 집단화하기 위해 앞 절에서 소개한 유사 클래스를 구성하고 있는 키워드들 사이의 유사 정도를 (식 5)를 적용하여 문서들 사이의 의미적 연결 정도를 측정할 수 있도록 제안한다.

$$\mu_{D_i} = \frac{1}{n}\sum_{i=1}^{n}\mu_C(k_i) \text{----(식 5)}$$

$n$:유사 클래스를 구성하는 키워드의 개수

$\mu_C(k_i)$: 유사 클래스를 구성하는 키워드의 유사정도

$\mu_{D_i}$ : 문서( $D_i$)에 대한 유사 클래스 $C_i$의 유사정도

또한, 의미적으로 유사한 내용을 가지는 문서들을 분류하기 위해 임의의 $\alpha$값을 추출하여 $\alpha$-cut을 적용한다. 예를 들어 0.5-cut을 적용하여 생성된 유사 클래스를 나타내는 위의 그림 7에서 임의의 유사 클래스( $C_i$)를 구성하고 있는 키워드 $w_1, w_2, w_3$에 대해 문서에서의 소속 정도를 나타내면 표 13과 같다.

표 13 호환 클래스의 키워드에 대한 소속 정도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .99 | 0 | 0 | .10 | .40 | 1 | .65 | 0 | .40 | 0 |
| 2 | .40 | .40 | 1 | 0 | .90 | .20 | 1 | .90 | .99 | .10 |
| 3 | 0 | 1 | .10 | .95 | .65 | 1 | .99 | 0 | .94 | .94 |

따라서, (식 5)에 적용하여 각 문서에 대한 호환 클래스의 유사 정도는 표 14와 같고 0.6-cut을 적용했을 때 구성되는 문서 분류 집합은 $D=\{d_5, d_6, d_7, d_9\}$으로 구성된다.

표 14 문서 집합에서 문서의 소속 정도

| D \ C | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $D_i$ | .46 | .46 | .37 | .35 | .65 | .73 | .88 | .30 | .78 | .37 |

결과적으로 특정 도메인의 문서들에 대해 각 문서들이 가지는 의미적 유사 정도의 값을 키워드의 발생 빈도에서 유도하였고 동일한 단어나 단어구가 반복적으로 출현하는 문서들에 대해서는 서로 의미적으로 연결된다고 볼 수 있어 같은 문서 집합으로 분류하게 된다.

### 4.4 질의어 확장과 문서 분류 알고리즘

본 절은 문서 집합에서 추출한 키워드의 빈도를 이용하여 사용자의 관심도를 보다 적절하게 표현할 수 있도록 질의어를 확장하고 의미적으로 연결되는 문서들에 대해서는 동일한 카테고리를 생성하는 문서 분류 알고리즘을 제안한다.

임의의 문서에서 특정 키워드의 발생 빈도가 높으면 높을수록 그 키워드는 문서에서 중요한 의미를 가지고 문서의 주제를 많이 반영한다고 할 수 있으므로 적절한 키워드를 추출하는 것이 선행되어야 한다.

본 논문에서는 형태소 분석기를 이용하여 추출한 키워드의 빈도를 이용하였고 제안하는 질의어 확장과 문서 분류 알고리즘은 퍼지 논리에 기반을 둔 퍼지 집합간의 관계성을 정의하는데 역점을 두었다.

먼저, 질의어를 확장시키는 알고리즘은 문서 집합에서 추출한 각각의 키워드들에 대하여 퍼지 관계성을 적용하여 관련성이 많은 키워드 집합을 생성함으로써 사용자 관심도를 보다 많이 표현하고, 또한 키워드 직접 매칭에 의한 검색 기법의 단점인 의미상으로 연결된 문서에 대한 검색을 해결한다.

또한, 문서간의 유사 정도에 따라 문서를 분류하는 알고리즘은 임의의 문서 결과 집합에서 생성된 키워드의 유사 클래스 이용하여 동의어 문제를 해결할 수 있도록 문서간의 유사 정도에 따라 문서를 분류한다.

따라서, 제안한 질의어 확장과 문서 분류 알고리즘을 적용한 검색 기법은 사용자의 관심도를 효율적으로 반영하고 의미적으로 관련 있는 문서를 동일한 카테고리에 배정함으로써 검색 시간을 줄일 수 있다.

질의어 확장과 문서 분류를 위한 알고리즘은 다음과 같다.

① 키워드의 발생 영역 정보와 빈도를 계산
입력 : 각 문서
출력 : 각 키워드 위치정보와 빈도

```
Procedure occurrence_frequency( )
{
  for (i=0 ; i<=n ; i++ ) {
    Preprocessing( Di )
    // 형태소 분석기를 사용하여 용어 분리
    for(j=0 ;  ; j++) {
      Extract_keyword( Kij)
      // 문서 i에서 키워드 Kj 을 추출
      Search_keyword_location( Ki )
      // 추출한 키워드의 위치 정보
      Occurrence_frequent ( Kj,  Di )
      // 문서 Di에서 추출된 키워드 Kj의 빈도 계산
```

```
    }
  }
}
```

② 키워드의 발생 위치에 따른 소속 정도 계산
입력: 각 키워드의 위치정보 및 빈도
출력: 각 키워드의 소속 정도

```
Procedure fuzzy_keywordrelation( Ki , Kj )
{
  Sigmoid_function(frequent_ ki)
  //시그모이드 함수를 적용, 소속 정도로의 변환
   {
    if(keyword_location ==(Title or keywordSet))
    //키워드의 위치에 따라 각기 다른 소속함수정의
    //키워드가 타이틀이나 키워드 집합에 발생
      Select_Sigmoid( S1)
     //시그모이드 S1 소속 함수를 선택하여 적용
      Occurrence_frequency( )
     //키워드의 빈도 계산
      Create_fuzzness(frequent)
     //빈도에 의한 소속 정도로의 사상
  else if (keyword_location==(Abstract or Conclusion))
   //키워드가 요약이나 결론부분에 발생
     Select_Sigmoid ( S2)
     //시그모이드 S2 소속 함수를 선택하여 적용
     Occurrence_frequency( )
     //키워드의 빈도 계산
     Create_fuzzness(frequent)
     //빈도에 의한 소속 정도로의 사상
  else (keyword_Location == Text)
   //키워드가 본문에 발생
     Select_Sigmoid( S3)
     // 시그모이드 S3 소속 함수를 선택하여 적용
     Occurrence_frequency( )
     // 키워드의 빈도 계산
     Create_fuzzness(frequent)
     // 빈도에 의한 소속 정도로의 사상
 }
for(i=0 ;   ; i++) {
   Binary_relation( ki, ki+1)
   // 키워드 사이의 관계성 생성
 }
}
```

③ 키워드들 간의 유사 클래스를 생성
입력: 각 키워드의 소속 정도
출력: 키워들 간의 유사클래스

```
Procedure extend_query( )
{
  Extract_ αValue(  )
  // α-cut을 수행하기 위한 임의의 α값 생성
  for(i=0;   ;i++) {
    if(membership_degree( kt) >= α)
   // α값 보다 큰 소속 값을 가진 키워드에 대하여
      Create_similarityclass ( ki)
     // 유사 클래스 생성
 }
```

④ 문서들간의 유사 정도 계산
입력: 키워드 유사클래스
출력: 문서들간의 유사도

```
Procedure fuzzy_documentrelation(  )
{
  Procedure Fuzzy_keywordrelation( Ki , Kj )
  // 키워드들 사이의 관계성 생성
  Procedure extend_query( Ck )
  // 유사 클래스 생성에 의한 질의어 확장
  for (i=0 ;   i++)
       Binary_relation( di,  di+1)
       // 문서사이의 유사도 계산
       Extract_ αValue ( )
       // α-cut을 수행하기 위한 임의의 α값 생성
}
```

⑤ 문서 카테고리 생성
입력: 각 문서의 소속정도
출력: 호환 클래스

```
Procedure document_categorization( )
{
   for (i=0;  ; i++) {
      if(membership_degree( di) >= α)
     // α값 보다 소속 값을 가진 문서에 대하여
       Create_compatibilityclass( di)
       // 호환 클래스 생성
   }
}
```

## 5. 실험 및 평가

본 장에서는 사용자의 관심도를 표현한 질의어에 대한 결과 집합의 문서를 대상으로 본 논문에서 제안하는 질의어 확장과 문서 분류 알고리즘을 적용하여 사용자 질의어와 유사도가 높은 키워드의 집합을 생성하여 의미적으로 연결된 문서끼리 집단화가 구성되는지를 알아본다.

이를 위해 'Altavista' 검색 엔진을 이용했을 때, 입력한 질의어에 대하여 탐색 문서의 결과 집합을 대상으로 한 실험을 한다. 사용자의 관심도를 표현하는 질의어 입력에 '인공지능'이라는 단일어(single word)를 입력했을 때 탐색된 200개의 웹(web) 문서로 구성된 결과 집합에서 임의의 문서 15개를 선택한다. 그리고 웹 상에 공개되어 있는 'HAM' 형태소 분석기[10]를 이용하여 15개의 문서에서 추출한 400개의 키워드 중 임의의 10개의 키워드를 선택한다.

다음은 각각의 문서와 키워드들로 구성된 집합을 나타낸다.

$D=\{D_1, D_2, D_3, \cdots, D_{15}\}$

$W=\{w_1, w_2, w_3, \cdots, w_{10}\}$

={인공지능, 신경망, 로보틱스, 클러스터링, 퍼지, 자동제어, 학습 규칙, 연결 가중치, 알고리즘, 뉴런}

위와 같은 실험대상을 이용하여 사용자의 관심을 표현한 질의어를 확장시키고 의미적으로 관련성이 있는 문서끼리 분류하기 위해 본 논문에서 제안한 알고리즘을 적용한다.

**[Step 1] 문서에서 추출한 키워드의 유사도를 계산한다.**

형태소 분석기를 이용하여 추출된 키워드들의 각 문서에서 발생 빈도를 계산한 값은 표 15와 같다.

$W=\{w_1, w_2, w_3, \cdots, w_{10}\}$

={인공지능, 신경망, 로보틱스, 클러스터링, 퍼지, 자동제어, 학습 규칙, 연결 가중치, 알고리즘, 뉴런}

표 15 문서에서의 키워드 빈도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8 | 2 | 3 | 0 | 2 | 3 | 7 | 3 | 4 | 2 | 7 | 0 | 2 | 9 | 3 |
| 2 | 7 | 3 | 0 | 3 | 0 | 3 | 7 | 8 | 2 | 0 | 1 | 8 | 6 | 9 | 2 |
| 3 | 7 | 2 | 4 | 0 | 12 | 5 | 4 | 9 | 2 | 9 | 0 | 1 | 12 | 1 | 9 |
| 4 | 0 | 6 | 2 | 12 | 2 | 12 | 2 | 0 | 2 | 0 | 1 | 0 | 1 | 8 | 1 |
| 5 | 2 | 3 | 7 | 9 | 1 | 4 | 1 | 1 | 3 | 1 | 1 | 0 | 6 | 10 | 0 |
| 6 | 7 | 13 | 3 | 2 | 10 | 2 | 0 | 7 | 16 | 13 | 9 | 0 | 10 | 1 | 10 |
| 7 | 3 | 3 | 9 | 2 | 9 | 10 | 2 | 9 | 13 | 12 | 0 | 9 | 13 | 1 | 12 |
| 8 | 1 | 7 | 2 | 2 | 8 | 2 | 1 | 1 | 2 | 8 | 12 | 3 | 11 | 2 | 0 |
| 9 | 4 | 8 | 9 | 1 | 9 | 9 | 8 | 1 | 2 | 0 | 10 | 18 | 2 | 0 | 1 |
| 10 | 5 | 1 | 2 | 12 | 0 | 2 | 7 | 2 | 4 | 2 | 15 | 12 | 3 | 1 | 2 |

추출된 키워드 빈도를 문서에서의 가능성 값으로 사상시키기 위해 시그모이드 함수를 이용한다. 이 때, 각 키워드에 가중치를 부여하기 위해 키워드의 출현 위치 정보와 빈도를 고려한 각각의 시그모이드 함수를 이용하였다. 그리고 각 문서에서의 임의의 키워드의 평균 중요 정도를 구하기 위해 (식 1)을 적용하였고 그 결과 값은 표 16과 같다.

표 16 가중치를 부여한 키워드의 중요 정도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | .90 | .40 | 0 | .70 | .85 | .99 | .40 | .50 | .20 | .99 | 0 | .20 | 1 | .40 |
| 2 | .99 | .40 | 0 | .40 | 0 | .85 | .99 | 1 | .90 | 0 | .80 | 1 | .94 | 1 | .70 |
| 3 | .99 | .20 | .50 | 0 | 1 | .80 | .50 | 1 | .70 | 1 | 0 | .10 | 1 | .40 | 1 |
| 4 | 0 | .94 | .70 | 1 | .90 | 1 | .70 | 0 | .20 | 0 | .10 | 0 | .10 | 1 | .80 |
| 5 | .90 | .40 | .99 | 1 | .10 | .50 | .10 | .80 | .40 | .80 | .10 | 0 | .94 | 1 | 0 |
| 6 | .99 | 1 | .40 | .20 | 1 | .20 | 0 | .99 | 1 | 1 | 1 | 0 | 1 | .40 | 1 |
| 7 | .40 | .85 | 1 | .70 | 1 | 1 | .20 | .10 | 1 | 1 | 0 | 1 | 1 | .10 | 1 |
| 8 | .10 | .99 | .20 | .90 | 1 | .20 | .10 | .10 | .20 | 1 | 1 | .85 | 1 | .20 | 0 |
| 9 | .95 | 1 | .10 | .10 | 1 | 1 | 1 | .40 | .20 | 0 | 1 | 1 | .90 | 0 | .10 |
| 10 | .80 | .10 | .70 | 1 | 0 | .70 | .99 | .20 | .50 | .70 | 1 | 1 | .40 | 0.1 | .20 |

**[Step 2] 키워드 유사 클래스를 이용한 질의어를 확장한다.**

질의어를 확장하기 위해, 먼저 (식 4)를 적용하여 다음 표 17과 같이 키워드들 사이의 유사 정도를 구하고, $\alpha$-cut을 적용하여 유사 클래스를 생성한다.

이러한 유사 클래스는 사용자가 입력한 질의어와 유사 정도가 일정한 이상한 값으로 구성되어 결과적으로 질의어를 확장할 수 있게 한다.

표 17 키워드들 사이의 유사 정도

| W \ W | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | .59 | .51 | .58 | .49 | .55 | .38 | .41 | .64 | .47 |
| 2 | .59 | 1 | .54 | .40 | .48 | .50 | .41 | .36 | .62 | .56 |
| 3 | .51 | .54 | 1 | .42 | .54 | .61 | .57 | .45 | .56 | .45 |
| 4 | .58 | .40 | .42 | 1 | .45 | .31 | .52 | .46 | .45 | .41 |
| 5 | .49 | .48 | .54 | .45 | 1 | .40 | .45 | .45 | .37 | .52 |
| 6 | .55 | .50 | .61 | .31 | .40 | 1 | .58 | .58 | .52 | .36 |
| 7 | .38 | .41 | .57 | .52 | .45 | .58 | 1 | .61 | .50 | .49 |
| 8 | .41 | .36 | .45 | .46 | .45 | .58 | .61 | 1 | .61 | .52 |
| 9 | .64 | .62 | .56 | .45 | .37 | .52 | .50 | .61 | 1 | .57 |
| 10 | .47 | .56 | .45 | .41 | .52 | .36 | .49 | .52 | .57 | 1 |

본 실험에서는 키워드들 사이의 유사 정도가 0.6 이상인 키워드들로 구성된 유사 클래스를 생성한다. 이에 대한 그림은 다음과 같다.

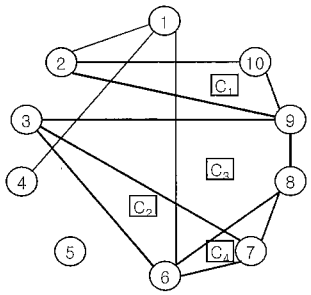


그림 10 0.6-cut을 적용한 유사 클래스 생성

따라서, 문서 결과 집합에서 추출된 10개의 키워드들의 유사 관계에 의해 생성된 유사클래스는 다음과 같다.

$C_1 = \{w_2, w_9, w_{10}\} = \{$신경망, 알고리즘, 뉴런$\}$

$C_2 = \{w_3, w_6, w_7\} = \{$로보틱스, 자동제어, 학습규칙$\}$

$C_3 = \{w_3, w_7, w_8, w_9\} = \{$로보틱스, 학습규칙, 연결가중치, 알고리즘$\}$

$C_4 = \{w_6, w_7, w_8\} = \{$자동제어, 학습규칙, 연결가중치$\}$

**[Step 3] 문서들 간의 카테고리를 생성한다.**

문서 카테고리를 생성하기 위해서는 유사 클래스를 구성하는 키워드와 문서의 유사도 값에 (식 4)을 이용하여 임의의 문서 카테고리에 문서가 속할 가능성 값을 생성하고 $\alpha$-cut을 적용한다.

다음은 각각의 키워드의 유사 클래스에 따라 문서 분류의 생성과정을 보여준다.

① $C_1 = \{w_2, w_9, w_{10}\} = \{$신경망, 알고리즘, 뉴런$\}$일 경우 $C_1$ 유사 클래스를 구성의 키워드와 문서간의 유사도 값은 표 18과 같다.

표 18 $C_1$를 구성하는 키워드와 문서간의 유사도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | .99 | .40 | 0 | .40 | 0 | .85 | .99 | 1 | .90 | 0 | .80 | 1 | .94 | 1 | .70 |
| 9 | .95 | 1 | .10 | .10 | 1 | 1 | 1 | .40 | .20 | 0 | 1 | 1 | .90 | 0 | .10 |
| 10 | .80 | .10 | .70 | 1 | 0 | .70 | .99 | .20 | .50 | .70 | 1 | 1 | .40 | 0.1 | .20 |

표 18의 값에 (식 4)를 적용하면, 표 19와 같이 임의의 문서 카테고리( $D_i$)에 분류될 가능 정도의 값이 생성된다.

표 19 문서 카테고리( $D_1$)에 속할 정도 값

| C \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $D_i$ | .91 | .50 | .26 | .50 | .33 | .85 | .99 | .53 | .53 | .23 | .93 | 1 | .47 | .33 | .33 |

따라서, 0.6-cut을 적용할 경우 문서 카테고리 ( $D_1$)에 속할 문서는 $\{d_1, d_6, d_7, d_{11}, d_{12}\}$이다.

② $C_2 = \{w_3, w_6, w_7\} = \{$로보틱스, 자동제어, 학습규칙$\}$일 경우 $C_2$유사 클래스를 구성의 키워드와 문서간의 유사도 값은 표 20과 같고, 표 21은 임의의 문서 카테고리( $D_i$)에 분류될 가능 정도의 값을 나타낸다.

표 20 $C_2$를 구성하는 키워드와 문서간의 유사도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | .99 | .20 | .58 | 0 | 1 | .80 | .50 | 1 | .70 | 1 | 0 | .10 | 1 | .40 | 1 |
| 6 | .99 | 1 | .48 | .20 | 1 | 20 | 0 | .99 | 1 | 1 | 1 | 0 | 1 | .40 | 1 |
| 7 | .40 | .85 | 1 | .70 | 1 | 1 | .20 | .10 | 1 | 1 | 0 | 1 | 1 | .10 | 1 |

표 21 문서 카테고리( $D_2$)에 속할 정도 값

| C \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $D_i$ | .79 | .68 | .68 | .30 | 1 | .66 | .23 | .69 | .90 | .1 | .33 | .36 | 1 | .30 | 1 |

여기서 0.6-cut을 적용하면

$D_2 = \{d_1, d_2, d_3, d_5, d_6, d_8, d_9, d_{10}, d_{13}, d_{15}\}$이다.

③ $C_3 = \{w_3, w_7, w_8, w_9\} = \{$로보틱스, 학습규칙, 연결가중치, 알고리즘$\}$

일 경우 $C_3$유사 클래스를 구성하는 키워드와 문서간의 유사도 값은 표 22이다.

표 22 $C_3$를 구성하는 키워드와 문서간의 유사도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | .99 | .20 | .50 | 0 | 1 | .80 | .50 | 1 | .70 | 1 | 0 | .10 | 1 | .40 | 1 |
| 7 | .40 | .85 | 1 | .70 | 1 | 1 | .20 | .10 | 1 | 1 | 0 | 1 | 1 | .10 | 1 |
| 8 | .10 | .99 | .20 | .90 | 1 | .20 | .10 | .10 | .20 | 1 | 1 | .85 | 1 | .20 | 0 |
| 9 | .95 | 1 | .10 | .10 | 1 | 1 | 1 | .40 | .20 | 0 | 1 | 1 | .90 | 0 | .10 |

따라서, 임의의 문서 카테고리( $D_i$)에 분류될 가능 정도의 값은 표 23이다.

표 23 문서 카테고리( $D_3$)에 속할 정도 값

| C \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $D_i$ | .61 | .76 | .45 | .42 | 1 | .75 | .45 | .40 | .52 | .75 | .50 | .73 | .97 | .17 | .52 |

여기서 0.6-cut을 적용하면

$D_3 = \{d_1, d_2, d_5, d_6, d_{10}, d_{12}, d_{13}\}$이다.

④ $C_4 = \{w_6, w_7, w_8\} = \{$자동제어, 학습규칙, 연결가중치$\}$

일 때, $C_4$유사 클래스를 구성의 키워드와 문서간의 유사

도 값은 표 24이다.

표 24 $C_4$를 구성하는 키워드와 문서간의 유사도

| W \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | .99 | 1 | .40 | .20 | 1 | .20 | 0 | .99 | 1 | 1 | 1 | 0 | 1 | .40 | 1 |
| 7 | .40 | .85 | 1 | .70 | 1 | 1 | .20 | .10 | 1 | 1 | 0 | 1 | 1 | .10 | 1 |
| 8 | .10 | .99 | .20 | .90 | 1 | .20 | .10 | .10 | .20 | 1 | 1 | .85 | 1 | .20 | 0 |

따라서, 임의의 문서 카테고리($D_i$)에 분류될 가능 정도의 값은 표 25이다.

표 25 문서 카테고리($D_4$)에 속할 정도 값

| C \ D | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $C_i$ | .49 | .94 | .53 | .60 | 1 | .46 | .10 | .39 | .73 | 1 | .66 | .61 | 1 | .23 | .66 |

여기서, 0.6-cut을 적용하면

$D_4 = \{d_2, d_4, d_5, d_9, d_{10}, d_{11}, d_{12}, d_{13}, d_{15}\}$이다.

따라서, 유사 클래스를 구성하는 키워드에 따라 각기 다른 문서를 가지는 문서 분류가 이루어짐을 볼 수 있고 이것을 그림으로 나타내면 그림 11과 같다.

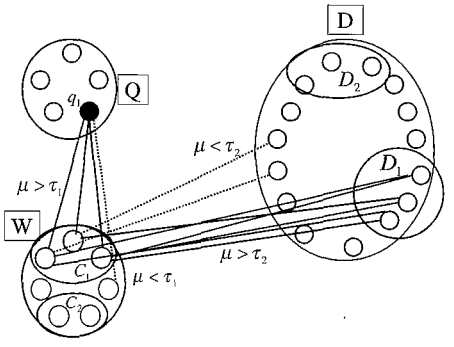


그림 11 유사클래스에 따른 문서의 카테고리

위 그림은 사용자가 입력한 질의어와 문서에서 추출한 색인어 사이의 소속정도가 $\alpha$값인 $\tau_1$보다 큰 값으로 구성되는 유사클래스를 생성하여 질의어의 확장과 유사클래스를 구성하는 색인어와 문서간의 소속 정도가 $\tau_2$보다 큰 값으로 구성되는 문서에 대한 카테고리의 생성을 보여주고 있다.

본 논문에서 퍼지 관계성을 이용하여 특정 도메인에(본 논문에서는 '인공지능')관한 문서를 대상으로 질의어 확장과 문서분류 알고리즘을 적용하여 실험한 결과, 문서를 대표하는 색인어의 발생 빈도가 문서의 주제별 분류에 매우 중요한 식별 기능을 가지고 있음을 알 수 있어 효율적으로 문서를 분류할 수 있는 기법을 제공할 수 있다.

## 6. 결론 및 향후 연구과제

본 논문은 각 문서의 내용을 대표할 수 있는 키워드를 추출하여 사용자의 관심도를 반영하는 질의어를 확장하고, 또한 의미적으로 연결되는 문서끼리 집단화를 구성하는 문서 분류 알고리즘을 제안한다.

제안한 질의어 확장과 문서 분류 알고리즘은 기존의 기법과는 다르게 문서에서 추출한 키워드의 발생 빈도를 이용하여 키워드가 해당 문서에서의 중요 정도를 나타내기 위해 퍼지 소속 함수를 정의하고 키워드와 키워드 사이의 관계와 문서간의 관계를 퍼지 관계로 정의하였다.

또한, 키워드의 가중치를 부여하기 위하여 추출한 키워드의 발생 영역에 따라 각기 다른 소속함수를 정의하여 사용자의 관심도를 보다 많이 반영하도록 하였다.

질의어 확장 기법은 각 문서에서 발생하는 키워드의 소속 정도의 값이 퍼지 유사관계를 만족할 때 $\alpha$-cut을 수행하여 유사도가 높은 키워드, 즉 의미적으로 연결되는 키워드들 사이의 유사 클래스를 생성하여 질의어 확장을 수행한다.

문서 분류는 질의어 확장을 위해 생성된 각각의 유사 클래스의 키워드들에 대하여 문서간의 유사 정도를 계산하고 유사도가 높은 문서끼리 문서 분류를 수행하여 의미적으로 연결되는 문서에 대하여 같은 문서 집단으로 형성시킨다.

따라서, 본 논문에서 제안한 질의어 확장과 문서분류의 알고리즘을 적용한 검색 기법은 사용자의 관심도를 많이 반영한 문서를 선별, 제공하며 의미적으로 연결된 문서를 동일한 문서 집단으로 분류함으로써 단순히 키워드 직접 매칭에 의한 검색 기법보다 사용자에게 적합한 문서를 제공한다.

향후 연구과제로는 제안한 문서 분류 알고리즘을 적용하여 생성된 문서 집단에서 자동적으로 색인화 할 수 있는 기법과 내용적으로 연결된 문서들에 대해 계층적으로 분류하여 문서들을 체계적인 분류할 수 있는 기법이며, 이를 통해 임의의 문서가 데이타베이스에 저장되었을 때 의미적으로 연관성이 많은 문서 집단으로 자동 분류시키며 문서 집합을 대표할 수 있는 색인어를 이용하여 유사한 문서들의 검색 속도를 높이고 정확률과 재현율을 향상시키도록 한다.

## 참 고 문 헌


[1] 이종득,"시소러스 기반의 정보검색 시스템 구축을 위한 개념 그룹화 방법", 전북대학교 대학원 박사학위 논문, 1998. 2.

[2] R. Baeza-ates, B. Ribeiro-Neto, "Modern Information Retrieval," p.230-255, 1998.

[3] P. Wallis, J. A. Tom, "Relevance judgements for assessing recall," Information Processing and Management 32, pp. 273-286, 1998.

[4] 조광제, 김준태, "역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류", 정보과학회학술발표논문집, 1996.10.

[5] 정영미,"정보검색론", 구미무역사, 1997.

[6] G. J. Klir, B. Yuan, "Fuzzy Sets and Fuzzy Logic Theory and Applications," 1998.

[7] L. T. Kóczy, "Information retrieval by fuzzy relations and hierarchical co-occurrence," 1997

[8] P. Baranyi, T. D. Gedeon, L. T. Kóczy, "Improved fuzzy and neural network algorithms for frequency prediction in document filtering," TR 97-02, 1997.

[9] L. T. Kóczy, T. D. Gedeon, J. A. Kóczy, "The construction of fuzzy relational maps in information retrieval," IETR 98-01, 1998.

[10] 강승식, 이하규, "한국어 형태소 분석기 HAM의 형태소 분석 및 철자 검사 기능", 한글 및 한국어 정보처리학회 학술발표논문집, 1998.

[11] Ricardo Baeza-Yates, Betthier Ribeiro-Neto, "Modern Information Retrieval," 1999.

[12] L. T. Kóczy, T. Gedeon, "Information retrieval by fuzzy relations and hierarchical co-occurrence," Part I. TR97-01, Dept. of Info. Eng., School of Comp. Sci. & Eng., UNSW, 1997.

[13] M. Blosseville, G. Hebrail, M. Monteil, N. Penot. "Automatic document classification: natural language processing, statistical analysis, and expert system techniques used together," SIGIR' 97. 1997

[14] P. Jacobs, "Using statistical methods to improve knowledge-based news categorization," IEEE Expert, 1998.

[15] R. Hoch "Using Information Retrieval techniques for text classification in document analysis," SIGIR' 98, 1998.

[16] 하얀, 최봉진, 김용성, 김순기, "2단계 필터링을 이용한 문서 선별 및 순위", 한국정보과학회 봄 학술 발표논문집(B) 제26권 제1호, 1999.

[17] 최봉진, 하얀, 황용주, 김용성, "Fuzzy Logic을 기반으로 한 SDI 서비스 설계", 한국정보과학회 가을 학술발표논문집( I ), 제25권, 제2호, 1998.

[18] 최동시, 정경택, "카테고리와 키워드의 밀접성 정보에 의한 문서 자동 분류 시스템 설계 및 구현", 정보과학회 학술발표논문집, 1995.



은 희 주

1998년 2월 전북대학교 컴퓨터과학과 졸업(이학사). 2000년 2월 전북대학교 전산통계학과 졸업(이학석사). 2000년 3월 ~ 현재 전북대학교 컴퓨터통계정보학과 박사과정. 관심분야는 퍼지논리, 정보검색, 인공지능, 전자 도서관 등



하 얀

1992년 2월 덕성여자대학교 전산학과 졸업(이학사). 1994년 8월 이화여자대학교 교육대학원 전자계산교육전공 졸업(교육학 석사). 2000년 2월 전북대학교 대학원 전산통계학과 졸업(이학박사). 2000년 9월 ~ 2001년 2월 중앙대학교 정보통신연구소 연구전담교수. 2001년 3월 ~ 현재 경인여자대학 멀티미디어 정보전산학부 전임강사. 관심분야는 XML 응용, 객체지향 모델링, 컴포넌트 모델링, 애니메이션 컴포넌트, 전자도서관, 퍼지 정보검색

김 용 성

정보과학회논문지 : 소프트웨어 및 응용 제 28 권 제 1 호 참조