단어 / 단어쌍 특징과 신경망을 이용한 두 문서간 유사도 측정

Measurement of Document Similarity using Term / Term - pair Features and Neural Network

저자 (Authors)
김혜숙, 박상철, 김수형
Hye Sook Kim, Sang Cheol Park, Soo Hyung Kim

출처 (Source)
정보과학회논문지 : 소프트웨어 및 응용 31(12), 2004.12, 1660-1671(12 pages)
Journal of KISS : Software and Applications 31(12), 2004.12, 1660-1671(12 pages)

발행처 (Publisher)
한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE00617746

APA Style
김혜숙, 박상철, 김수형 (2004). 단어 / 단어쌍 특징과 신경망을 이용한 두 문서간 유사도 측정. 정보과학회논문지 : 소프트웨어 및 응용, 31(12), 1660-1671

단어/단어쌍 특징과 신경망을 이용한 두 문서간 유사도 측정

(Measurement of Document Similarity using Term/Term-pair Features and Neural Network)

김 혜 숙 † 박 상 철 † 김 수 형 ††
(Hye Sook Kim) (Sang Cheol Park) (Soo Hyung Kim)

요 약 본 논문은 두 문서간 유사도 측정 방법을 제안한다. 제안한 유사도 측정 모델의 주안점은 문서간 관련성의 정도를 두 문서간 일치하는 단어(term)및 단어쌍(term-phrase)에 기반하여 이들이 해당 문서에서 차지하는 가중치를 통해 측정하는 것이다. 유사도 측정 과정에 영향을 미치는 특징을 설계함에 있어 기존의 연구들이 하나의 특징만을 고려하였던 것에 비하여 본 논문은 여러 가지 특징들을 고려한다. 즉, 단어뿐만 아니라 단어쌍과 관련된 특징을 결합하여 신경망을 통해 유사도를 측정한다. 제안된 방법의 우수성을 입증하기 위해 두 가지 측면에서 실험하였다. 첫 번째는 두 문서의 동일성 여부를 검증하는 문제이며, 두 번째는 다수의 문서를 대상으로 유사한 문서를 찾는 검색 문제이다. 이 두 가지 실험 모두에서 제안 방법이 기존의 Cosine 유사도 계산 방법 및 구색인 방법에 비해 우수한 성능을 보였다.

키워드 : 문서간 유사도, 신경망, 단어 빈도수, 구색인

Abstract This paper proposes a method for measuring document similarity between two documents. One of the most significant ideas of the method is to estimate the degree of similarity between two documents based on the frequencies of terms and term-pair, existing in both the two documents. In contrast to conventional methods which takes only one feature into account, the proposed method considers several features at the same time and meatures the similarity using a neural network. To prove the superiority of our method, two experiments have been conducted. One is to verify whether the two input documents are from the same document or not. The other is a problem of information retrieval with a document as the query against a large number of documents. In both the two experiments, the proposed method shows higher accuracy than two conventional methods, Cosine similarity measurement and a term-pair method.

Key words : Document Similarity, Neural Network, Term-Frequency, Term-pair Indexing

1. 서 론

인터넷의 정보와 다양한 서비스가 빠른 속도로 증가하고 있으며 이를 만들고 사용하는 사람의 수 또한 증가하고 있다. 따라서 대용량의 정보를 보다 빠르게 분류하고 효율적으로 관리하며 쉽게 검색할 수 있는 방안이 모색되어야 한다[1]. 아울러 많은 양의 문서를 관리하고 이를 효율적으로 검색하기 위한 문서 분류 모델에 관한 연구도 오래 전부터 계속되고 있다[2].

문서 클러스터링 및 분류 분야에서 문서간 관련성을 정량적으로 측정하기 위해 문서간의 유사도를 계산해야 한다. 문서간 유사도 측정을 통한 문서 분류는 분류 시간을 단축시키고 분류 효율(efficienccy)을 향상시키는데 기여할 수 있다. 그러나 지금까지 발표된 대부분의 유사도 측정 방법들은 대용량의 문서집합(corpus)을 대상으로 문서에 출현한 단어 빈도수 만을 사용하기 때문에 웹 상에 존재하는 다양한 형태의 문서들을 처리하는데 적용하기는 어렵다[3]. 문서간 유사도 측정기술의 응용분야로는 정보검색, 문서동일성 여부 검증, 문서분류 및 clustering등이 있다[4].

본 논문에서는 과거의 연구들을 바탕으로 임의의 두

· 본 연구는 한국과학재단 목적기초연구(R05-2003-10396-0) 지원으로 수행되었음
† 비 회 원 : 전남대학교 전자계산학과
hsfight@hanmail.net
sanchun@chonnam.ac.kr
†† 종신회원 : 전남대학교 전자컴퓨터정보통신공학부 교수
shkim@chonnam.ac.kr
논문접수 : 2003년 11월 27일
심사완료 : 2004년 10월 5일

문서간 유사도를 측정할 수 있는 새로운 모델을 제안한다. 특히 단어색인 방법과 구색인 방법에서 사용되는 특징들을 결합하여 문서간 유사도 측정의 특징으로 삼는다. 이러한 특징을 바탕으로 신경망을 통해 문서간 유사도를 정량적인 수치로 산출한다. 본 논문은 여러 가지 측면에서 다양한 유사도 계산 모델을 함께 적용하는데 주안점을 두고 연구를 진행하였다. 제안된 유사도 측정 방법의 성능은 두 문서간 동일성 여부의 검증 및 대용량의 문서를 대상으로 한 문서 검색 측면에서 살펴본다. 기존의 문서간 유사도 측정 방법인 Cosine 유사도 계산 방법 및 구색인 방법과 비교한 결과 제안 방법이 향상된 성능을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 정보검색 분야에서 하나의 질의를 대상으로 여러 개의 문서간 유사도 비교에 일반적으로 사용되는 Cosine 유사도 계산 방법에 대해 살펴보고 구색인 방법에서 색인어가 되는 공기의 의미와 여러 가지 관련 연구에 대해 살펴본 후, 3장에서는 제안 방법의 문서간 유사도 측정에 대해 설명하며, 4장에서는 이를 검증하기 위한 실험 및 결과에 대해 기술한다. 마지막으로 5장에서 결론 및 향후 연구를 제시한다.

2. 관련연구

본 장에서는 단어색인 방법을 통한 Cosine 유사도 계산 방법과 구색인에 의한 유사도 측정 방법에 대하여 설명한다.

2.1 Cosine 유사도

문서간 유사도 측정을 위한 색인추출 방법 중 하나는 개별 단어로 구성된 색인을 생성하는 방법이다. 단어색인은 단일단어로 구성되어 적은 양의 데이타에서도 많은 색인을 추출할 수 있다는 장점이 있는 반면에 문맥정보를 포함할 수 없다는 단점이 있다[5,6].

색인어는 의미적으로 해당 문서의 주제를 기억하는데 도움을 주는 단어이다. 일반적으로 색인어는 명사로 국한하게 되는데, 이는 명사가 자체 의미를 지니고 있어서 구분하거나 이해하기 쉽기 때문이다. 형용사나 부사, 접속사 등은 주로 보조적이기 때문에 색인어로는 덜 유용하다[3]. 추출된 임의의 색인어 k가 해당 문서에서 어느 정도의 중요도를 갖는지를 평가하기 위해 가중치 $w(k)$를 부여하게 되는데, 표 1은 지금까지 알려진 다양한 가중치 공식[2,7]을 보여주고 있다. 여기에 사용되는 변수는 다음과 같다.

d_i : $i-$번째 문서

$K_i = \{ k_{ij} \mid 1 \le j \le n_i \}$: 문서 d_i 내의 모든 단어 색인어 집합

n_i : $i-$번째 문서 d_i 내의 색인어 개수

$f(k_{ij})$: 문서 d_i 에서 색인어 k_{ij}의 출현 빈도수(*frequency*)

$w(k_{ij})$: 문서 d_i 에서 색인어 k_{ij}의 가중치(*weight*)

$f(k_{ij}, k_{ir})$: 문서 d_i 에서 두 색인어 k_{ij}와 k_{ir}이 동시에 출현한 빈도수

$P_i = \{ (k_{ij}, k_{ir}) \mid 1 \le j, r \le n_i \}$: 단어쌍 색인어 집합

표 1 색인어 k_{ij}에 대한 가중치 $w(k_{ij})$계산 공식

이 름	단어빈도에 따른 가중치공식
이진 함수	$w(k_{ij}) = \begin{cases} 1, & \text{if } f(k_{ij}) > 1 \\ 0, & otherwise \end{cases}$
단순 함수	$w(k_{ij}) = f(k_{ij})$
로그 함수	$w(k_{ij}) = 1 + \log f(k_{ij})$
더블로그 함수	$w(k_{ij}) = 1 + \log(1 + \log f(k_{ij}))$
루트 함수	$w(k_{ij}) = \sqrt{f(k_{ij})}$
보정 함수	$w(k_{ij}) = (1-\alpha) + \alpha \times \frac{f(k_{ij})}{\max f(k_{ij})}$, 단 $\alpha = 0.5$ 또는 0.6
오가피 함수	$w(k_{ij}) = \frac{f(k_{ij})}{2 + f(k_{ij})}$
더블로그2 함수	$w(k_{ij}) = 1 + \log_2(1 + \log_2 f(k_{ij}))$
루트직선 함수	$w(k_{ij}) = \frac{f(k_{ij}) + 3}{4}$

표 1의 가중치공식[8]을 바탕으로 만들어진 유사도 측정 모델중 하나인 벡터 모델은 다음과 같다. 문서 d_i내의 색인어 k_{ij}의 가중치 $w(k_{ij})$는 양의 실수이며, 질의 색인어도 가중치를 가진다. 가중치 계산에 사용되는 공식은 표 1의 어떤 것이라도 무방하다. 즉, 질의 벡터 $\vec{q}$는 $\vec{q} = \langle w(k_{q1}), w(k_{q2}), \cdots, w(k_{qt}) \rangle$로 정의되며, 문서 d_i는 벡터 $\vec{d}_i = \langle w(k_{i1}), w(k_{i2}), \cdots, w(k_{it}) \rangle$ 로 표현된다. 여기서 질의 q의 색인어 집합은 $K_q = \{ k_{qj} \mid 1 \le j \le n_q \}$ 로 가정하고, 문서 d_i의 색인어 집합은 $K_i = \{ k_{ij} \mid 1 \le j \le n_i \}$ 로 가정했을 때, $K = K_q \cup K_i$이며 $t = |K|$이다.

문서 d_i와 사용자 질의 q는 그림 1과 같이 t 차원 벡터로 표시된다. 벡터 모델에서 문서 d_i와 질의 q의

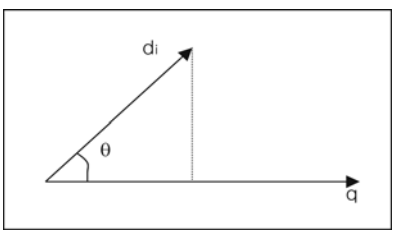

그림 1 코사인 Θ 값으로 $Sim(d_i, q)$측정

유사도 측정은 두 벡터 $\vec{d_i}$와 $\vec{q}$의 상관도로 구할 수 있으며, 이 상관도의 예로 두 벡터간 사이각의 코사인 값으로 정량화 할 수 있다[8,9].

결론적으로 문서 d_i와 질의 q의 Cosine 유사도는 식 (1)과 같이 나타낼 수 있다.

$$Sim(d_i, q) = \frac{\vec{d_i} \cdot \vec{q}}{|\vec{d_i}| \times |\vec{q}|} = \frac{\sum_{a=1}^{t} w(k_{ia}) \times w(k_{qa})}{\sqrt{\sum_{a=1}^{t} w(k_{ia})^2} \times \sqrt{\sum_{a=1}^{t} w(k_{qa})^2}} \quad (1)$$

식 (1)에서 $|\vec{d_i}|$와 $|\vec{q}|$는 문서와 질의 벡터의 norm 값으로, $|\vec{q}|$요소는 모든 문서에 동일하며, $|\vec{d_i}|$는 유사도 값을 정규화 하기위한 요소이다. $w(k_{ia})$와 $w(k_{qa})$가 0보다 크거나 같은 값을 갖기 때문에 문서와 질의간 유사도 값은 0과 1사이의 값이 된다. 즉, 벡터 모델은 질의 q와 문서 d_i내에 공통적으로 존재하는 정보의 양을 기준으로 유사도를 측정하기 때문에 부분적으로 질의에 정합 되는 문서는 모두 0 이 아닌 유사도 값을 갖는다. 유사도 값은 공통된 색인의 수가 같을 때 각 문서에서 색인의 수가 적을수록 높은 유사도를 갖으며 긴 문서에서 많은 색인이 꼽히면 그만큼 공통 색인의 수가 나올 확률이 높기에 이런 경우 불이익을 주기 위한 것이다.

Cosine 유사도는 유사도 측정의 가장 대중적인 모델로 단순하고 빠르기에 용어 가중치 기법이 검색성능을 향상시키고 부분정합으로 질의에 근접한 문서검색이 가능하며 이를 순위화 해준다. 단점으로는 색인어의 상호 독립성 가정이있다. 즉 색인어 k_{ij}의 가중치 $w(k_{ij})$와 색인어 k_{ir}의 가중치 $w(k_{ir})$는 서로에게 영향을 끼치지 않는다.

벡터 모델에서 문서 d_i와 질의 q의 유사도 측정은 두 벡터 $\vec{d_i}$와 $\vec{q}$의 상관도로 구하였다. 따라서 문서 d_x와 문서 d_y의 유사도 측정은 $\vec{d_x}$ 와 $\vec{d_y}$의 상관도로 구할 수 있다. 즉, 두 문서 d_x와 d_y를 각 단어 색인어의 가중치로 구성된 벡터로 표현하면 $\vec{d}_x = \langle w(k_{x1}), w(k_{x2}), \cdots, w(k_{xt}) \rangle$, $\vec{d}_y = \langle w(k_{y1}), w(k_{y2}), \cdots, w(k_{yt}) \rangle$로 나타낼 수 있으며, 두 문서간 유사도 계산식은 식 (2)로 정의된다. 여기서 t값은 $|K_x \cup K_y|$이다.

$$Sim(d_x, d_y) = \frac{\vec{d_x} \cdot \vec{d_y}}{|\vec{d_x}| \times |\vec{d_y}|} = \frac{\sum_{a=1}^{t} w(k_{xa}) \times w(k_{ya})}{\sqrt{\sum_{a=1}^{t} w(k_{xa})^2} \times \sqrt{\sum_{a=1}^{t} w(k_{ya})^2}} \quad (2)$$

2.2 구색인에 의한 문서간 유사도 계산 모델

구색인 방법[10,11]은 공기 정보 등을 추출하기 위해 단어쌍을 색인으로 추출한다. 공기(collocation 또는 co-occurrence)란 두 단어가 동일 문서, 동일 문단, 동일 문장 또는 일정한 크기의 단어창 안에서 같이 발생하는 현상을 말하며 공기 빈도수가 클수록 두 단어가 밀접한 관련이 있다고 간주한다[12-15]. 즉, 임의의 두 단어가 일정영역에서 동시에 출현하는 빈도가 클수록 두 단어의 관련이 깊다는 것이다[16].

문서 d_i에서 임의의 단어쌍 (k_{ij}, k_{ir})의 가중치 $w(k_{ij}, k_{ir})$를 계산하기 위해 해당 단어쌍의 출현 빈도수 $f(k_{ij}, k_{ir})$와 단어쌍의 정보량 $INFO(k_{ij}, k_{ir})$을 고려하는데, 단어쌍의 정보량은 식 (3)과 같이 정의된다[5].

$$INFO(k_{ij}, k_{ir}) = -\log_2 P(k_{ij}, k_{ir}) \approx -\log_2 (P(k_{ij}) \times P(k_{ir})) \quad (3)$$

$P(k_{ij})$와 $P(k_{ir})$는 색인어 k_{ij}와 k_{ir}이 문서 d_i내에서 출현할 확률을 의미하며 $P(k_{ij}) = f(k_{ij}) / \sum_{a=1}^{n_i} f(k_{ia})$, $P(k_{ir}) = f(k_{ir}) / \sum_{a=1}^{n_i} f(k_{ia})$과 같이 계산할 수 있다. 한편 단어쌍 (k_{ij}, k_{ir})의 가중치 $w(k_{ij}, k_{ir})$은 식 (4)에서와 같이 단어쌍 (k_{ij}, k_{ir})이 문서내에서 출현한 빈도수 $f(k_{ij}, k_{ir})$에 정보량 $INFO(k_{ij}, k_{ir})$을 곱한 값으로 정의된다.

$$w(k_{ij}, k_{ir}) = f(k_{ij}, k_{ir}) \times INFO(k_{ij}, k_{ir}) \quad (4)$$

식 (4)의 가중치를 문서의 크기에 관계없이 상대적 비교가 가능하도록 하기 위해 식 (5)와 같이 정규화한다[2].

$$u_i(k_{ij}, k_{ir}) = \frac{(w(k_{ij}, k_{ir}) - \overline{w_i})}{\sigma_i}$$

$$\text{단, } \overline{w_i} = \frac{\sum_{j,r=1}^{n_i} w(k_{ij}, k_{ir})}{n_i(n_i - 1)/2},$$

$$\sigma_i = \sqrt{\frac{\sum_{j,r=1}^{n_i} (w(k_{ij}, k_{ir}) - \overline{w_i})^2}{n_i(n_i - 1)/2}} \quad (5)$$

여기서 n_i는 문서 d_i내의 전체 색인어 수이다. 식 (5)와 같이 표준화된 단어쌍의 가중치는 문서간 유사도를 측정하는 기준으로 사용될 수 있다. 즉, 단어쌍을 색인어를 바탕으로 문서 d_x와 d_y간 유사도를 계산하는 식은 다음과 같다.

$$Sim(X, Y) = \sum_{(v_1, v_2) \in P_x \cap P_y} (u_x(v_1, v_2) \times u_y(v_1, v_2)) \quad (6)$$

식 (6)에서 P_x와 P_y는 문서 d_x와 d_y에 나타난 단어쌍 색인어 집합이며, $P_x \cap P_y$는 문서 d_x와 d_y에 공통으로 출현하는 단어쌍의 집합을 나타낸다. 또한

$u_x(v_1, v_2)$와 $u_y(v_1, v_2)$는 문서 d_x와 d_y에 공통으로 출현하는 단어쌍에 대한 정규화된 가중치를 의미한다. 따라서, 단어쌍을 기반으로 계산되는 유사도는 두 문서의 색인 파일에 공통으로 존재하는 단어쌍의 가중치 값의 곱들을 합산한 것으로 이 값이 클수록 두 문서간 유사도가 높아진다. 단어쌍을 색인어로 하는 구색인 방법은 문맥의 정보를 어느정도 반영할 수 있기에 단어색인 방법에 비해 문서의 내용을 더 잘 내포할 수 있다. 하지만 색인어가 되는 단어쌍 추출에 드는 시간적인 비용이 크다[5,13,17].

3. 문서간 유사도 측정

3.1 제안 방법의 시스템 구조

본 논문에서 제안하는 유사도 측정 시스템은 그림 2와 같이 색인 추출부, 특징 추출부, 유사도 측정부로 구성되어 있다.

첫째, 색인 추출부는 단어 색인어 집합과 단어쌍 색인어 집합을 추출한다. 단어 색인의 경우 형태소 분석을 통해 추출된 명사만을 색인어로 사용하였으며, 구(단어쌍)색인의 경우는 추출된 모든 명사들을 대상으로 동일 문장의 범위 내에서 단어쌍을 조합하여 색인어로 추출한다. 둘째, 특징 추출 부분에서는 색인 추출부에서 얻어진 결과를 이용하여 유사도 계산을 위한 여섯 가지 특징들을 추출한다. 셋째, 유사도 측정 부분에서는 다층 신경망에 6가지 특징 값을 적용하여 두 문서간 유사도를 계산한다.

3.2 색인 생성

색인이란 문서의 내용을 분석하여 그 문서를 다른 문서들로부터 구별할 수 있도록 하는 단어(term) 또는 단어쌍(term-pair)을 의미한다. 실제로 한국어에서 문서 분리도가 높은 단어들은 주로 개념을 표현하는 명사와 고유명사에 밀집되어 있기에 단어색인 방법에서 유사도 측정을 위한 색인어는 명사만을 대상으로 한다. 구색인 방법에서도 명사만을 대상으로 단어쌍을 형성하여 이를 색인어로 추출한다.

구색인 방법에서 인접한 단어 사이의 공기 정보를 추출하기 위해 슬라이딩 윈도우 기법[5]을 사용하는데, 먼저 추출된 내용어의 순서열에 일정 크기의 윈도우(window)를 설정한 후 윈도우의 맨 앞의 내용어와 다음 내용어들 간의 쌍을 추출한다(윈도우의 크기는 10으로 설정). 윈도우는 문장의 처음에서부터 마지막 내용어까지 움직이되, 문장의 끝에서 문장의 경계를 넘지 않도록 한다[16,18,19]. 윈도우가 문장의 경계를 넘지 않게 하는 이유는 서로 다른 문장에 속해 있는 내용어가 같은 문장내의 내용어 보다 약한 문맥 정보를 가지기 때문이며, 아울러 추출되는 색인의 수를 일정 수준으로 유지하기 위해서이다. 그림 3은 본 논문에서 사용한 슬라이딩 윈도우 기법을 이용하여 색인어를 추출한 예이다.

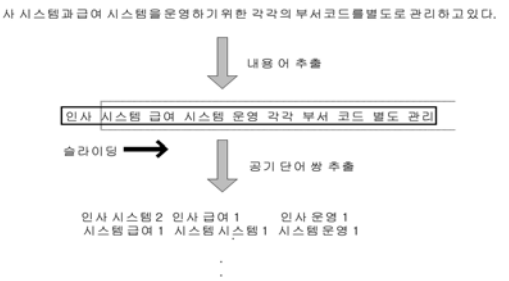

그림 3 슬라이딩 윈도우 기법 예

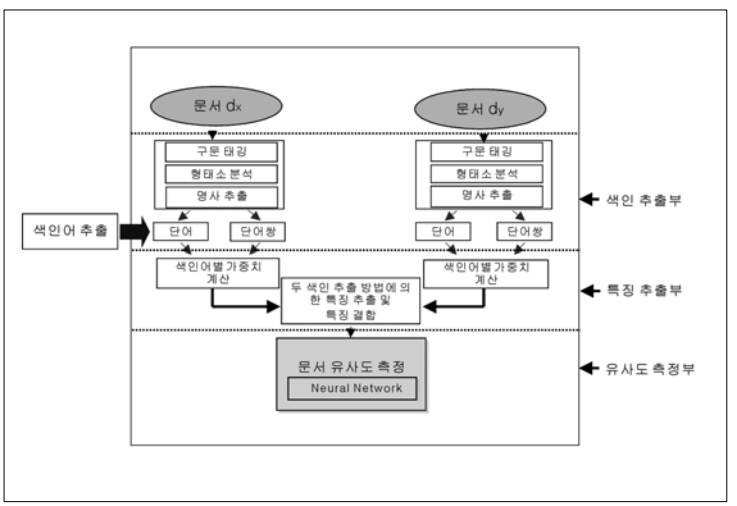

그림 2 제안된 유사도 측정 시스템 구조

3.3 특징 추출

본 논문에서는 두 문서간 유사도 측정을 위해 여섯 가지 특징을 사용한다. 이중 다섯 가지는 단어색인 방법을 기반으로 설계된 것이고, 나머지 한 가지는 구색인 방법을 바탕으로 만들어진 것이다. 단어색인 방법의 특징으로는 두 문서간 일치하는 색인어의 수, 상호정보량, 평균조건확률, 색인어 출현 빈도를 바탕으로 한 전형적인 Cosine 유사도, 최소 가중치의 합 등 다섯 가지이다. 여섯 번째 특징은 구색인 방법에서 사용하는 두 문서간 일치하는 단어쌍에 대한 가중치의 합을 이용한다.

(1) 두 문서간 일치하는 색인어의 수

두 문서간 유사도 측정시 기초가 되는 자료로 일치하는 색인어수를 우선순위로 고려한다. 이 특징은 단순하여 계산에 소모되는 시간적인 비용이 거의 없는 반면, 문서의 문맥적인 구조까지 고려된 유사도 측정에는 한계점이 있다.

$$F_1(d_x, d_y) = |K_x \cap K_y| \quad (7)$$

여기서 K_x, K_y는 문서 d_x와 d_y의 색인어 집합이며, $K_x \cap K_y$는 두 문서에 공통으로 출현하는 색인어의 집합이다. 따라서 식 (7)의 F_1값은 두 문서에 공통으로 출현하는 색인어의 수를 의미한다.

(2) 상호정보량

상호정보(Mutual Information)는 통계적 언어 모델(statistical language model)에서 단어간의 연관관계 측정에 이용되는 개념으로, 두 단어 중 한 단어가 출현했다는 사건이 다른 단어의 출현 여부 예측에 기여하는 정도를 수치적으로 나타낸 값이다[20]. 문서 d_i내의 전체 색인어수가 n_i일 때, 색인어 k_{ij}와 색인어 k_{ir}이 출현한 빈도수를 각각 $f(k_{ij})$와 $f(k_{ir})$라고 하고 특정 크기 윈도우에서 동시에 출현한 빈도수를 $f(k_{ij}, k_{ir})$ 라고 했을 때 n_i가 충분히 크면 상호정보량 $MI(k_{ij}, k_{ir})$는 다음과 같이 근사될 수 있다[2].

$$MI(k_{ij}, k_{ir}) \approx \log_2 \frac{n_i \times f(k_{ij}, k_{ir})}{f(k_{ij}) \times f(k_{ir})} \quad (8)$$

위 식 (8)에 의해 계산된 상호정보 $MI(k_{ij}, k_{ir})$는 색인어 k_{ij}와 색인어 k_{ir}사이의 통계적 상관관계를 나타내는데 $MI(k_{ij}, k_{ir})$의 값이 커짐에 따라 색인어 k_{ij}와 색인어 k_{ir}사이의 상관 관계의 긴밀성은 증가한다.

본 논문에서는 두 문서 d_x와 d_y간 유사도 측정을 위한 두 번째 특징을 식 (9)에서와 같이 각 문서의 색인어 출현 빈도수를 바탕으로 계산한다.

$$F_2(d_x, d_y) = \frac{\sum_v f_x(v) \times f_y(v)}{\sum_{a=1}^{n_x} f_x(k_{xa}) + \sum_{a=1}^{n_y} f_y(k_{ya})}$$

단, $v \in K_x \cap K_y$ (9)

여기서 n_x, n_y는 각각 문서 d_x와 d_y의 색인어 개수를 의미한다. 식 (8)은 두 색인어 사이의 상호정보인 반면 식 (9)는 식 (8)을 두 문서 사이의 상호정보량을 나타낼수 있도록 확장한 것이다. 식 (8)에서 두 색인어가 동시에 출현한 빈도수가 크면 두 색인어 사이의 연관성이 높게 계산되듯이, 식 (9)에서도 두 문서간에 공통으로 출현하는 색인어가 많을수록 상호연관성이 높아진다. 즉, $F_2(d_x, d_y)$는 $F_1(d_x, d_y)$보다 두 문서간의 연관성을 좀더 구체적으로 산출하는 특징이라 할 수 있다. 이 특징은 검색성능 향상에 유용하다. 하지만 저빈도 단어사이의 상호정보량이 고빈도 단어사이의 상호정보량보다 상대적으로 과대평가될 수 있다.

(3) 평균조건확률

평균조건확률(Average Conditional Probability)은 두 조건부확률 $P(k_{ij} \mid k_{ir})$와 $P(k_{ir} \mid k_{ij})$의 평균값이다. 조건확률 $P(k_{ij} \mid k_{ir})$는 색인어 k_{ij}의 빈도수 $f(k_{ij})$와 관계없이 색인어 k_{ij}과 k_{ir}와의 공기 빈도수 $f(k_{ij}, k_{ir})$가 크면 그 값이 커지며 색인어 k_{ij}가 색인어 k_{ir}에 얼마나 종속적인가를 나타내고 있다. 반대로 $P(k_{ir} \mid k_{ij})$는 색인어 k_{ir}의 색인어 k_{ij}에 대한 의존도를 나타낸다. 이 두 가지 조건부확률은 두 색인어의 공기 빈도수가 각 색인어의 빈도수 크기에 따라 중요도가 다르다는 전제에서 출발한다. 만일 색인어 k_{ij}가 색인어 k_{ir}에 비해 더 저빈도 색인어라면, 두 색인어의 조건확률은 저빈도 색인어 k_{ij}에 더 큰 영향을 받게 된다. 따라서 정규화하기 위해 두 조건확률의 평균값을 취한다[21,22].

$$SA(k_{ij}, k_{ir}) = \frac{P(k_{ij} \mid k_{ir}) + P(k_{ir} \mid k_{ij})}{2} = \frac{1}{2}\left(\frac{P(k_{ij}, k_{ir})}{P(k_{ir})} + \frac{P(k_{ij}, k_{ir})}{P(k_{ij})}\right) \approx \frac{f(k_{ij}, k_{ir})}{2}\left(\frac{1}{f(k_{ij})} + \frac{1}{f(k_{ir})}\right) \quad (10)$$

본 논문에서는 식 (10)을 바탕으로 두 문서간 유사도 측정을 위한 세 번째 특징을 식(11)과 같이 계산한다.

$$F_3(d_x, d_y) = \frac{1}{2}\left(\frac{\sum_v f_x(v) \times f_y(v)}{\sum_{a=1}^{n_x} f_x(k_{xa})} + \frac{\sum_v f_x(v) \times f_y(v)}{\sum_{a=1}^{n_y} f_y(k_{ya})}\right)$$

단, $v \in K_x \cap K_y$ (11)

식 (10)은 두 색인어 사이의 평균조건확률인 반면 식 (11)은 식 (10)을 두 문서 사이의 평균조건확률을 나타낼 수 있도록 확장한 것이다.

(4) Cosine 유사도

두 문서 d_x와 d_y간 유사도 측정은 $\overrightarrow{d_x}$와 $\overrightarrow{d_y}$의 상관도로 하되, 각 단어에 대한 가중치는 표 1의 단순함수

인 $w(k) = f(k)$ 에 의해 계산한다. 이를 바탕으로 두 문서간 유사도 계산식은 식 (12)와 같다.

$$F_4(d_x, d_y) = \frac{\sum_{a=1}^{t} f_x(k_{xa}) \times f_y(k_{ya})}{\sqrt{\sum_{a=1}^{t} f_x(k_{xa})^2} \times \sqrt{\sum_{a=1}^{t} f_y(k_{ya})^2}} \quad (12)$$

(5) 최소 가중치의 합

최소 가중치 합은 유사도 측정에 중요한 척도로 간주되며, 정보검색 분야에서 대표적으로 사용되는 단어 가중치 계산 모델이다. 이는 한 문서에서 임의의 색인어 k에 대한 가중치는 해당 색인어의 빈도수를 전체 색인어의 총 빈도수로 나눈 값으로 간주한다. 이때 두 문서간 유사도는 식 (13)에서와 같이 두 문서에서 공통으로 출현하는 단어에 대한 두 가중치중 최소치를 취하며 이들을 합산한 값으로 정의할 수 있다.

$$F_5(d_x, d_y) = \sum_{v} \min\left(\frac{f_x(k_{xa})}{\sum_{a=1}^{n_x} f_x(k_{xa})}, \frac{f_y(k_{ya})}{\sum_{a=1}^{n_y} f_y(k_{ya})} \right)$$

단, $v \in K_x \cap K_y$ (13)

(6) 두 문서간 일치하는 단어쌍에 대한 가중치 합

단어쌍을 색인어로 하는 문서간 유사도는 문서 d_x와 d_y에서 공통으로 출현하는 단어쌍에 대한 정규화된 가중치의 곱들을 합산한 값이다. 2.2절에서 설명된 식 (6)을 바탕으로 두 문서 d_x, d_y사이의 유사도 계산식은 식 (14)와 같다.

$$F_6(d_x, d_y) = \sum_{(v_1, v_2) \in P_x \cap P_y} (u_x(v_1, v_2) \times u_y(v_1, v_2)) \quad (14)$$

$P_x \cap P_y$는 문서 d_x와 d_y에 공통으로 출현하는 단어쌍 색인어의 집합이며, $u_x(v_1, v_2)$와 $u_y(v_1, v_2)$는 문서 d_x 와 d_y 에 공통으로 출현하는 단어쌍에 대한 정규화된 가중치이다. 식 (14)는 두 문서간 유사도 측정에 있어 일치하는 단어쌍이 어느 정도 충분히 발생하지 않는 경우 유사도 측정에 큰 영향을 미치지는 못한다는 단점이 있다.

3.4 신경망을 이용한 유사도 측정

앞 절에서 추출된 특징들은 값의 범위가 서로 달라 이들을 신경망에 그대로 적용하는데는 각 특성의 반영면에서 문제가 있다. 실제로 본 연구에서 채택한 특징 중 f_1(두 문서간 일치하는 색인어의 수)의 값은 평균적으로 약 5 정도의 정수값을 갖는 반면, f_4(코사인유사도)는 0과 1사이의 실수값을 취하기 때문에 이들을 선형결합할 경우 f_1의 특성은 많이 반영되는 반면, f_4의 특성은 조금 밖에 반영되지 못하는 문제점이 있다. 따라서 본 논문에서는 신경망에 값을 제공하기 전에 특징값들을 정규화하는 함수를 추가하여 scalability문제를 줄였다. 정규화된 각각의 특징값들에 신경망을 적용하여 두 문서간 유사도를 측정한다.

제안 방법에 의한 문서간 유사도 측정에 고려되는 신경망 구조는 그림 4와 같다. 신경망 입력으로는 단어색인 방법과 구색인 방법에 의해 추출된 6개의 특징과 bias를 입력으로 하고, hidden layer는 bias를 제외하고 4개의 node를 갖는다. Output layer는 1개의 node를 갖는데, node의 출력값을 통해 두 문서간의 유사도를 정량적인 수치로서 측정 할 수 있다. 각 노드에서의 활성화 함수는 0과 1사이의 실수값을 생성하는 sigmoid 함수를 사용하였으며 오류역전파 학습 알고리즘을 사용하여 신경망의 연결가중치를 훈련하였다. 훈련 시 적용된 학습률은 0.15이며, 최대 반복횟수는 1000회이다. 두 문서간 유사도가 0.5보다 크면 두 문서간 관련성이 높고 그렇지 않으면 두 문서간 관련성이 낮다고 간주한다.

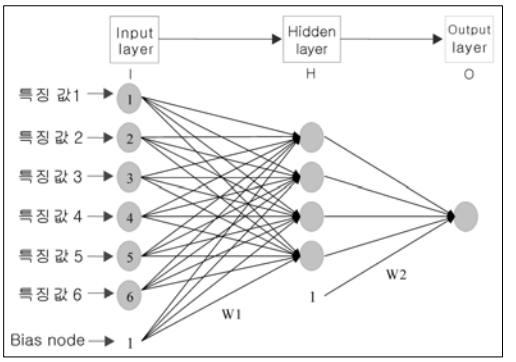

그림 4 문서유사도 측정 위한 신경망 구조도

신경망 대신 6차원 특징 $F=(f_1, f_2, \cdots, f_6)$을 선형결합한 $\alpha_1 f_1 + \alpha_2 f_2 + \cdots + \alpha_6 f_6$ 형태의 판별함수를 고려해 볼 수 있는데 이 방법은 두 가지 문제점이 있다. 첫째, 여섯 가지 특징 각각의 값의 범위(scale)가 서로 다르기 때문에 각 특징의 특성이 고루 반영되기 어렵고 둘째, 유사한 문서와 유사하지 않은 문서 두 부류 사이의 결정경계(decision boundary)가 비선형(nonlinear) 함수의 형태를 취하는 경우 효과적인 판별기준을 제공하지 못한다. 본 연구에서도 두 부류 사이의 결정경계가 선형이 아님을 실험을 통해 확인할 수 있었다. 따라서 서로 다른 scale의 특징들을 효과적으로 결합하고, 다차원 공간의 비선형 결정함수를 효과적으로 학습하기 위해 그림 4와 같은 신경망을 통해 두 문서간 유사도를 측정하였다.

4. 실험 및 평가

본 논문에서 제안한 유사도 측정방법의 신뢰성을 입증하기 위하여 100개의 문서를 대상으로 다음과 같은

두 가지의 실험을 수행하였다. 첫째는 임의의 두 문서에 대한 동일성 여부의 검증이고, 두 번째는 다수의 문서를 대상으로 유사한 문서를 검색하는 문제이다.

4.1 실험 데이타 구성

문서간 유사도 측정 실험을 위하여 사용한 컴퓨터는 700MHz의 속도를 갖는 Pentium III PC이고, 구현을 위하여 사용한 프로그래밍 언어는 Microsoft의 Visual C++이다. 실험문서는 작가, 주제 면에서 중복 없이 선정된 소설이나 수필을 대상으로 100개를 선정하였으며 각 문서의 크기는 A4 용지 1장 규격으로 통일하였다. 즉 100개의 문서 $D_1, D_2, \cdots, D_{100}$ 각각을 양분하여 $(D_{11}, D_{12}), (D_{21}, D_{22}), \cdots, (D_{100,1}, D_{100,2})$으로 분할한 후 $(D_{11}, D_{12}), \cdots, (D_{50,1}, D_{50,2})$는 훈련용으로 사용하고 $(D_{51,1}, D_{51,2}), \cdots, (D_{100,1}, D_{100,2})$는 테스트용으로 사용한다. 이때 각 문서는 그림 5에서와 같이 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, 9:1분할의 9가지 형태를 취했다. 각각의 분할에 대해 11개의 문서로 구성하였으며 9:1분할에 대하여만 12개의 문서로 구성하였다.

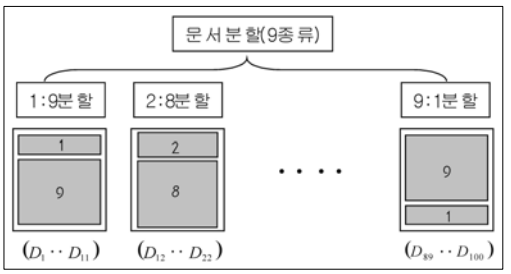

그림 5 문서간 유사도 측정 위한 실험 데이타 구성

4.2 문서 동일성 여부의 검증

4.2.1 실험 방법

문서 동일성 여부 검증이란 임의의 두 문서가 동일한 문서로부터 분할된 문서인지의 여부를 검증하는 것으로, 일종의 2-class 분류 문제이며 검증의 절차는 그림 6과 같다.

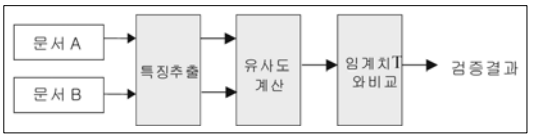

그림 6 문서 분류 절차

이 분류 시스템에서는 두 문서간에 계산된 유사도값을 이용하여 그림 7과 같이 오분류 확률이 최소가 되도록 분류한다. 여기서 오분류 확률이란 두 가지 경우 즉 FRR(동일한 문서인데 다른 문서라고 잘못 판단한 확률) 및 FAR(다른 문서인데 동일 문서라고 잘못 판단할 확률)을 포함한다. 그림 7에서 임계치(T)를 구하는 과정은 다음과 같다. 동일 문서쌍 사이의 유사도 값과 비동일 문서쌍 사이의 유사도 값 모두 정규분포를 이룬다고 가정하고, 동일 문서쌍 사이의 유사도 값들의 분포에 대한 평균 μ_A와 분산 σ_A, 비동일 문서쌍 사이의 유사도 값들의 분포에 대한 평균 μ_B과 분산 σ_B을 기준으로 식 (15)의 식에 의해 임계치(T)를 결정한다. 식 (15)에서 $P(A)=P(B)=0.5$라고 가정한다.

$$T=-2\ln(P(A)/\sigma_A)+(\frac{x-\mu_A}{\sigma_A})^2+2\ln(P(B)/\sigma_B)-(\frac{x-\mu_B}{\sigma_B})^2 \quad (15)$$

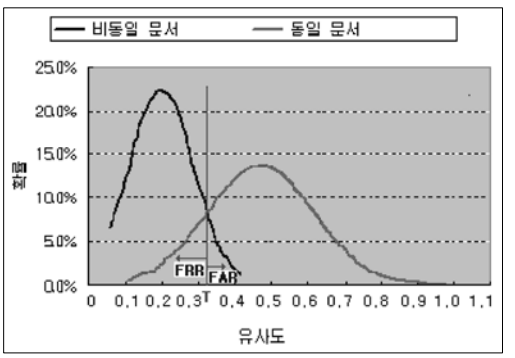

그림 7 동일 문서와 비동일 문서간의 유사도에 대한 오분류 확률 (FRR, FAR)

4.2.2 실험 데이타

동일한 문서쌍에 대한 유사도 분포(μ_A, σ_A)를 훈련하기 위해 50개의 문서쌍 $(D_{11}, D_{12}), \cdots, (D_{50,1}, D_{50,2})$을 사용하였으며, 비동일 문서쌍에 대한 유사도 분포(μ_B, σ_B)를 구하기 위해 50개의 문서쌍 $(D_{11}, D_{50,2})$, $(D_{21}, D_{49,2}), \cdots, (D_{50,1}, D_{12})$을 사용하였다. 제안 방법의 신경망 훈련을 위해 동일 문서쌍에 대한 목표 출력값은 1.0, 비동일 문서쌍에 대한 목표 출력값은 0을 부여하였다. 한편, 문서 검증 결과의 성능분석을 위해 50개의 동일 문서쌍 $(D_{51,1}, D_{51,2}), \cdots, (D_{100,1}, D_{100,2})$ 및 2,450개의 비동일 문서쌍 $(D_{i1}, D_{j2}), i, j=51, \cdots, 100 (i \neq j)$을 사용하였다.

4.2.3 실험 결과

임의의 두 문서의 동일성 여부를 검증하기 위해 먼저 두 문서사이의 유사도를 계산하고 계산된 유사도를 임계치와 비교하여 동일성 여부를 판단한다. 여기서는 세 가지의 유사도 측정 방법을 사용하였는데 표 2, 표 3,

표 4는 각각 Cosine 유사도 계산 방법, 구색인 방법에

표 2 Cosine 유사도 계산 방법을 통한 유사도 측정치

	$D_{51.1}$	$D_{52.1}$	$D_{53.1}$	$D_{54.1}$	$D_{55.1}$	$D_{56.1}$	$D_{57.1}$	$D_{58.1}$	$D_{59.1}$	$D_{60.1}$	$D_{61.1}$	$D_{62.1}$	$D_{63.1}$	$D_{64.1}$	$D_{65.1}$	$D_{66.1}$	…
$D_{51.2}$	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.07	
$D_{52.2}$	0.09	0.4	0.03	0.06	0.04	0.04	0.03	0.02	0	0.02	0.04	0.02	0.11	0.05	0.07	0.02	
$D_{53.2}$	0.01	0.01	0.14	0.03	0.1	0.02	0.03	0.11	0.02	0.05	0	0.12	0.06	0.16	0.04	0.06	
$D_{54.2}$	0.01	0.03	0	0.11	0	0.02	0	0	0	0	0	0.02	0.04	0	0	0	
$D_{55.2}$	0.09	0.03	0.06	0	0.12	0.06	0.02	0.05	0	0	0.05	0.02	0	0	0.03	0.07	
$D_{56.2}$	0.01	0.01	0.02	0.06	0.06	0.2	0	0.05	0.02	0.06	0.01	0.04	0.03	0.12	0.02	0.04	
$D_{57.2}$	0.04	0	0.01	0	0.07	0.01	0.32	0.01	0.01	0.03	0.02	0.02	0.06	0.02	0.03	0	
$D_{58.2}$	0.03	0.02	0.07	0.06	0.04	0.09	0.03	0.45	0.01	0.06	0	0.03	0.02	0.15	0.02	0.02	
$D_{59.2}$	0.09	0	0.05	0	0.04	0.02	0.02	0.02	0.2	0.01	0.03	0.08	0.09	0.02	0.05	0.04	
$D_{60.2}$	0.02	0	0.01	0.12	0	0.02	0.02	0.02	0	0.07	0.04	0.05	0	0.03	0	0.03	
$D_{61.2}$	0	0.11	0.05	0.02	0.03	0.01	0	0.04	0.02	0.08	0.46	0.05	0.02	0.19	0.03	0.04	
$D_{62.2}$	0.14	0.02	0.07	0.09	0.06	0.06	0.01	0.09	0.01	0.13	0.05	0.18	0.06	0.23	0.04	0.06	
$D_{63.2}$	0.34	0.14	0.09	0.07	0.13	0.03	0.01	0.04	0.04	0.03	0.05	0.13	0.39	0.03	0.17	0.07	
$D_{64.2}$	0.01	0.02	0.05	0.05	0.04	0.13	0.02	0.11	0.01	0.07	0.04	0.03	0.04	0.56	0.09	0	
:																	

표 3 구색인 방법을 통한 유사도 측정치

	$D_{51.1}$	$D_{52.1}$	$D_{53.1}$	$D_{54.1}$	$D_{55.1}$	$D_{56.1}$	$D_{57.1}$	$D_{58.1}$	$D_{59.1}$	$D_{60.1}$	$D_{61.1}$	$D_{62.1}$	$D_{63.1}$	$D_{64.1}$	$D_{65.1}$	$D_{66.1}$	…
$D_{51.2}$	4.49	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$D_{52.2}$	0.00	6.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$D_{53.2}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$D_{54.2}$	0.00	0.00	0.00	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$D_{55.2}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$D_{56.2}$	0.00	0.00	0.00	0.00	0.00	4.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$D_{57.2}$	0.00	0.00	0.00	0.00	0.00	0.00	1.89	0.00	2.06	1.98	0.00	0.00	0.00	0.00	0.00	0.00	
$D_{58.2}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$D_{59.2}$	0.00	1.03	0.00	0.00	0.00	0.00	0.00	0.00	1.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
$D_{60.2}$	0.00	0.00	0.00	0.00	0.76	0.00	0.00	0.00	1.02	3.45	0.00	0.00	1.02	0.00	0.00	0.00	
$D_{61.2}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	
$D_{62.2}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.01	0.00	0.00	0.00	0.00	
$D_{63.2}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	
$D_{64.2}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
:																	

표 4 제안 방법을 통한 유사도 측정치

	$D_{51.1}$	$D_{52.1}$	$D_{53.1}$	$D_{54.1}$	$D_{55.1}$	$D_{56.1}$	$D_{57.1}$	$D_{58.1}$	$D_{59.1}$	$D_{60.1}$	$D_{61.1}$	$D_{62.1}$	$D_{63.1}$	$D_{64.1}$	$D_{65.1}$	$D_{66.1}$	…
$D_{51.2}$	0.89	0.02	0.53	0.49	0.02	0.13	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.4	
$D_{52.2}$	0.68	1	0.13	0.26	0.06	0.16	0.07	0.05	0.02	0.05	0.07	0.05	0.57	0.15	0.24	0.07	
$D_{53.2}$	0.05	0.05	0.76	0.07	0.62	0.05	0.07	0.78	0.07	0.09	0.02	0.63	0.54	0.43	0.26	0.31	
$D_{54.2}$	0.11	0.14	0.13	0.99	0.02	0.1	0.02	0.12	0.02	0.13	0.02	0.07	0.14	0.02	0.02	0.02	
$D_{55.2}$	0.52	0.42	0.45	0.02	0.98	0.49	0.1	0.14	0.02	0.02	0.4	0.1	0.02	0.02	0.13	0.54	
$D_{56.2}$	0.03	0.03	0.03	0.07	0.07	0.97	0.02	0.06	0.05	0.14	0.03	0.06	0.08	0.11	0.04	0.07	
$D_{57.2}$	0.1	0.02	0.03	0.02	0.17	0.05	0.99	0.05	0.04	0.13	0.29	0.09	0.37	0.32	0.07	0.02	
$D_{58.2}$	0.17	0.09	0.42	0.24	0.22	0.54	0.04	0.85	0.04	0.12	0.02	0.08	0.08	0.53	0.1	0.04	
$D_{59.2}$	0.37	0.02	0.3	0.02	0.03	0.05	0.03	0.05	0.92	0.2	0.05	0.13	0.34	0.04	0.11	0.12	
$D_{60.2}$	0.06	0.02	0.25	0.2	0.02	0.05	0.1	0.05	0.02	0.89	0.05	0.18	0.02	0.11	0.02	0.13	
$D_{61.2}$	0.02	0.05	0.52	0.04	0.08	0.03	0.02	0.42	0.03	0.44	1	0.11	0.35	0.83	0.07	0.07	
$D_{62.2}$	0.22	0.03	0.08	0.14	0.33	0.05	0.03	0.07	0.02	0.11	0.05	1	0.09	0.12	0.05	0.08	
$D_{63.2}$	0.63	0.36	0.16	0.07	0.12	0.08	0.03	0.08	0.06	0.07	0.05	0.21	0.75	0.14	0.51	0.2	
$D_{64.2}$	0.05	0.1	0.05	0.17	0.22	0.86	0.05	0.74	0.05	0.18	0.25	0.09	0.22	1	0.43	0.02	
:																	

의한 유사도, 제안 방법에 의한 유사도 측정치의 일부를 발췌한 결과이다. 세 개의 표에서 공통적으로 관측할 수 있듯이 동일 문서쌍 ($D_{100,1}$, $D_{100,2}$), ($D_{51,1}$, $D_{51,2}$),…, ($D_{100,1}$, $D_{100,2}$)에 대한 유사도 값이 비동일 문서쌍 (D_{i1}, D_{j2}), $i,j=51,\cdots,100(i\neq j)$의 유사도 보다 대체로 높다.

표 5는 세가지 유사도 측정 방법 각각에 대한 임계치(T)를 보여준다. 즉 Cosine 유사도 계산 방법 및 구색인 방법의 경우 50개의 동일 문서쌍 및 50개의 비동일 문서쌍 각각에 대한 유사도 분포를 바탕으로 식 (16)에 의해 최적의 임계치를 계산하였으며, 제안 방법의 경우 신경망에 의해 계산되는 유사도 값이 0과 1사이로 정의되므로 그 중간값인 0.5를 임계치(T)로 선택하였다.

표 5 Cosine 유사도 계산 방법, 구색인 방법, 제안 방법을 통한 평균(μ), 분산(σ), 임계치(T)

문서 / 방법	동일 문서		비동일 문서		T
	μ_A	σ_A	μ_B	σ_B	
Cosine 유사도 계산 방법	0.3288	0.21804	0.0388	0.05509	0.141689
구색인 방법	2.68	3.0577	0.121	0.33102	0.846584
제안 방법	–	–	–	–	0.5

그림 8과 그림 9는 Cosine 유사도 계산 방법 및 구색인 방법 각각의 유사도 분포를 보여준다. Cosine 유사도 계산 방법의 경우 training data는 동일 문서의 경우 평균 0.3288과 분산 0.21804, 비동일 문서의 경우 평균 0.0388과 분산 0.05509를 갖는 정규분포를 따른다. 구색인 방법은 동일 문서의 경우 평균 2.68, 분산 3.0577, 비동일 문서의 경우 평균 0.121와 분산 0.33102를 갖는 분포를 따른다.

표 6은 표 5와 같이 구해진 임계값(T)을 기준으로 하여 각 방법 별로 test data(50개의 동일 문서쌍, 2,450개의 비동일 문서쌍)에 대한 정분류율과 오분류율을 나타

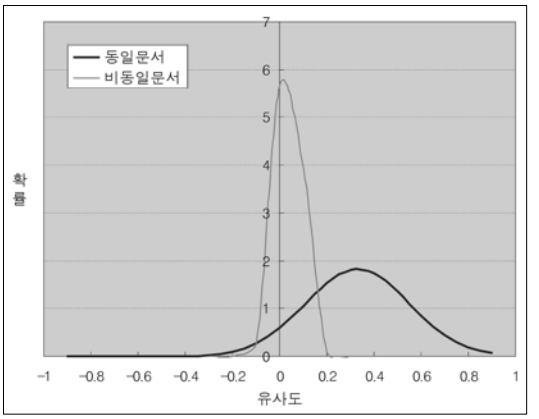

그림 8 Cosine 유사도 계산 방법을 통한 training data의 유사도 분포

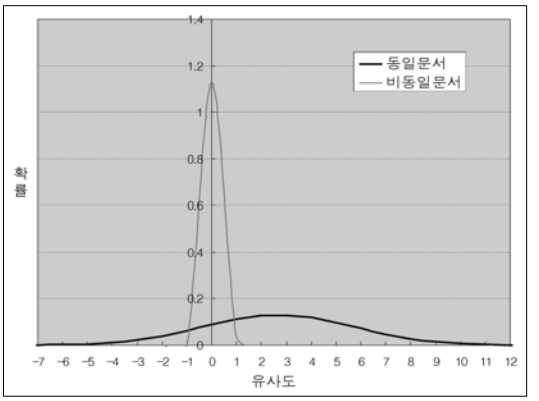

그림 9 구색인 방법을 통한 training data의 유사도 분포

표 6 세 가지 유사도 측정 방법에 따른 정분류율과 오분류율

		A (동일 문서)	B (비동일 문서)	계
A (동일 문서)	Cosine 유사도 계산 방법	76% =(38/50)	24% =(12/50)	50
	구색인 방법	70% =(35/50)	30% =(15/50)	
	제안 방법	**94%** =(47/50)	6% =(3/50)	
B (비동일 문서)	Cosine 유사도 계산 방법	5.51% =(135/2450)	94.49% =(2315/2450)	2,450
	구색인 방법	9.39% =(230/2450)	90.61% =(2220/2450)	
	제안 방법	9.51% =(233/2450)	**90.49%** =(2217/2450)	
계				2,500

낸 결과이다.

그림 10은 표 6의 결과를 도표로 나타낸 것이다. 표 6 또는 그림 10에서 알 수 있듯이 세 가지 유사도 측정 방법 중 제안 방법의 분류 성능이 가장 우수하다.

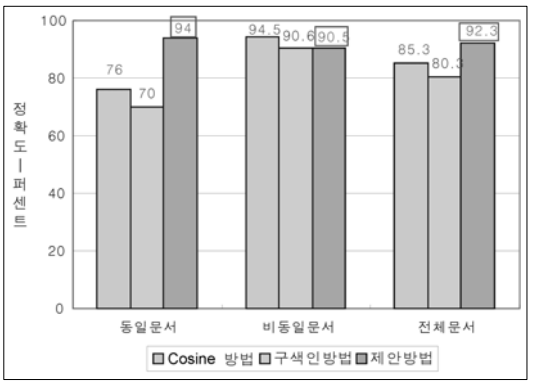

그림 10 세 가지 유사도 측정 방법에 따른 정분류율

위의 세 가지 유사도 측정방법에 의한 분류실험 이외에 추가로 사람들을 대상으로 문서 동일성 여부 검증에 대한 설문조사를 실시하였다. 설문 데이타 구성은 50개의 동일 문서쌍 ($D_{51,1}$, $D_{51,2}$),…, ($D_{100,1}$, $D_{100,2}$) 및 2,450개의 비동일 문서쌍 (D_{i1}, D_{j2}), $i,j=51,\cdots,100$ ($i\neq j$)에 대한 2,500개 test data중에서 동일 문서쌍 50개 ($D_{51,1}$, $D_{51,2}$),…,($D_{100,1}$, $D_{100,2}$)와 비동일 문서쌍 50개 ($D_{100,1}$, $D_{51,2}$),…, ($D_{51,1}$, $D_{100,2}$)등 총 100개의 문서쌍을 바탕으로 10가지 종류의 설문지를 구성하되, 각각의 설문지는 서로 다른 10개의 문서쌍을 포함하도록 하였다(그림 11 참조).

각각의 설문지별로 5명씩 총 50명을 대상으로 문서 동일성 여부에 대한 설문조사를 실시한 결과가 표 7에

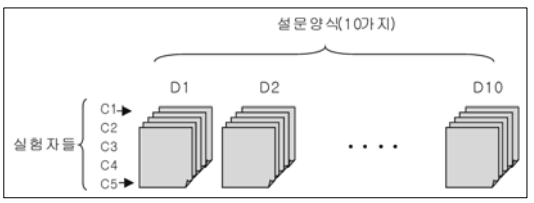

그림 11 사람을 대상으로 한 동일 문서 판정 여부 실험 데이타 구성

표 7 사람들 대상 동일 문서와 다른 문서의 분류

Input \ Output	A(동일 문서)	B(비동일 문서)
A	58%=(146/250)	42%=(104/250)
B	14%=(36/250)	86%=(214/250)

제시되어 있다. 표 7에서 알 수 있듯이 사람들은 비동일 문서를 비동일 문서라고 판정하는 경우가 동일 문서를 동일 문서로 판정하는 경우 보다 훨씬 우수한 성능을 보였지만 총 500개의 문항 중 360개에 대해서만 올바르게 응답하여 72%의 정확도를 보이는 사람들의 검증능력이 제안 방법 뿐 아니라 Cosine 유사도에 의한 방법 및 구색인에 의한 방법 보다 낮다는 것을 입증하고 있다.

표 8은 10가지 설문지 각각에 대하여 다섯명 사람들의 성능분포를 나타내고 있다. 이 표에서 알 수 있듯이 설문지별, 사람들 별 편차는 크지 않음을 알 수 있다.

4.3 문서 검색

4.3.1 실험 방법

문서 검색(retrival)은 50개의 기준 문서 $D_{51,1},\cdots,D_{100,1}$를 대상으로 $D_{51,2},\cdots,D_{100,2}$를 차례대로 제시하여 제시된 문서와 기준 문서들 사이의 유사도를 계산하고 이들 유사도 값을 바탕으로 기준 문서중 제시된 문서와 동일한 문서를 찾을 수 있는지 여부를 측정하는 실험이다(그림 12 참조). 검색의 정확도는 50개의 기준 문서를 제시된 문서와의 유사도를 바탕으로 정렬했을 때 제시된 문서와 동일한 기준문서가 상위 몇%안에 포함되는가로 나타낸다. 즉, 동일 문서 사이의 유사도가 비동일 문서간의 유사도에 비해 높은 값을 가질수록 정확한 검색을 했다고 할 수 있다.

그림 12는 실제 검색에서 사용된 데이타로 문서들을

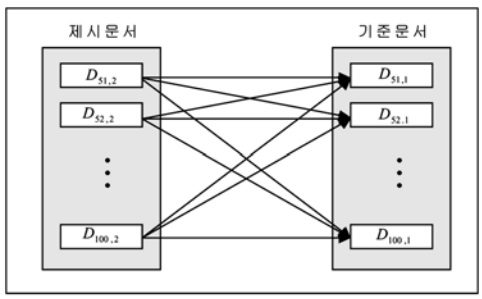

그림 12 임의의 입력문서에 대한 1:50 검색

표 8 사람들 대상 10개 설문양식에 대한 분포

설문양식	1	2	3	4	5	6	7	8	9	10	전체 평균
μ(평균)	7.2	7.8	7.4	7.8	6.8	6.2	8.2	7.0	6.2	7.4	7.2
σ(분산)	1.48	2.68	1.67	1.92	1.79	1.09	1.64	1.41	1.30	1.34	1.64

비교한 그림이다. 구체적인 예로, 제시문서에 해당하는 $D_{51,2}$문서와 기준문서 $D_{51,1}$부터 $D_{100,1}$의 문서 유사도를 계산한다. 그 후, 이 유사도를 바탕으로 50개의 기준 문서를 정렬 했을때 $D_{51,1}$이 1순위가 되면 가장 정확한 검색이라 할 수 있다. 본 검색 실험에서는 동일 문서가 상위 몇 %이내에 속하는 가를 2%, 4%, 10% 기준으로 살펴보았다.

4.3.2 실험 데이타

검색 실험의 경우에도 세 가지 유사도 측정 방법의 성능을 비교한다. Cosine 계산 방법 및 구색인 방법의 경우 그림 7에서와 같은 임계값(T) 계산이 필요하지 않으므로 별도의 훈련 데이타는 필요하지 않다. 제안 방법의 경우에는 유사도 계산을 위한 신경망 훈련을 위해 문서 동일성 여부의 검증실험에서와 같은 훈련 데이타를 사용한다. 즉, 50개의 동일 문서쌍 (D_{11}, D_{12}),…, ($D_{50,1}$, $D_{50,2}$) 및 50개의 비동일 문서쌍 (D_{11}, $D_{50,2}$), (D_{21}, $D_{49,2}$),…,($D_{50,1}$, D_{12})을 사용하였다.

4.3.3 실험 결과

표 9는 Cosine 유사도 계산 방법, 구색인 방법, 제안 방법에 대해 상위 2%, 4%, 10%기준의 검색 정확률을 나타낸다.

표 9 세 가지 유사도 측정 방법의 검색 정확률 비교

	Cosine 유사도 계산 방법		구색인 방법		제안 방법	
	개수	백분율	개수	백분율	개수	백분율
2%기준	33/50	66%	32/50	64%	40/50	80%
4%기준	37/50	74%	34/50	68%	45/50	90%
10%기준	42/50	84%	45/50	90%	47/50	94%

그림 13은 표 9의 결과를 도표로 나타낸 것으로 제안 방법이 다수의 문서를 대상으로 한 검색의 정확도 면에서 Cosine 유사도 계산 방법이나 구색인 방법보다 우수함을 알 수 있다.

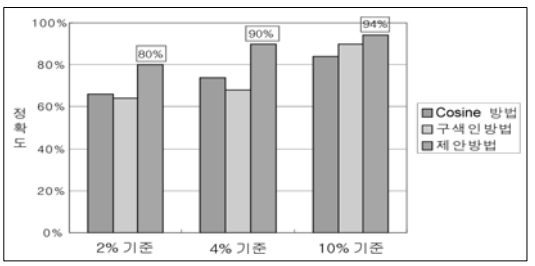

그림 13 세 가지 유사도 측정 방법에 대한 검색의 정확률 비교

5. 결론 및 향후 연구

본 논문에서는 두 문서간 유사도 측정 방법을 제안하였으며, 기존의 Cosine 유사도 계산 방법이나 구색인 방법과의 비교를 통해 본 연구의 우수성을 검증해보았다.

Cosine 유사도 계산 방법과 구색인 방법은 단어 색인어의 일반적인 정보인 빈도수(term-frequency)정보만을 이용하기 때문에 분류나 검색면에서 한계점이 드러난다. 따라서 본 논문에서는 단어색인 방법과 구색인 방법을 독립적으로 적용했을때 발생하는 한계점을 극복하고자 두 방법으로부터 유사도에 영향을 끼칠만한 여러 가지 특징들을 추출한 후, 신경망을 통해 두 문서간 유사도를 측정하였다. 이때 고려한 6가지 특징으로 두 문서간 일치하는 색인어 수, 상호정보량, 평균조건확률, Cosine 유사도, 최소 가중치 합, 두 문서간 일치하는 단어쌍에 대한 가중치 합 등을 고려하였다.

제안 방법의 우수성을 두 가지 측면에서 검증해 보았는데 첫 번째, 문서검증 면에서 Cosine 유사도 계산 방법과 구색인 방법, 제안 방법, 사람들을 대상으로 한 실험 등을 고려할 때 제안 방법이 가장 우수함을 확인할 수 있었다. 두 번째, 문서검색 면에서도 제안 방법이 Cosine 유사도 계산 방법이나 구색인 방법에 비해 16~22%정도 개선된 성능을 보임을 알 수 있었다.

향후 연구로는 문서간 유사도 측정에 있어 색인어에 기반 한 유사도 측정뿐 아니라, 문서에 내포되어 있는 의미까지 반영할 수 있도록 제안된 방법을 좀 더 개선하고 이에 대한 효과를 검증하는 것이다. 특히, 색인어 추출에 있어서 명사뿐 아니라 형용사, 부사, 접속사 등 문서 내에 쓰인 모든 단어들을 색인어로 고려해 보는 것도 흥미로운 일이 될 것이며, 단어 색인어의 가중치(중요도)를 결정하는 좀 더 다양한 방법을 모색할 필요성이 있다.

참 고 문 헌

[1] 허준희, 고수정, 김태용, 최준혁, 이정현, “문서의 주제어별 가중치와 말뭉치를 이용한 한국어 문서의 자동분류: 베이지안 분류자”, 한국정보과학회 가을 학술발표논문집, Vol. 26, No. 2, pp. 154-156, 1999.

[2] Y. Maarek, D. Berry and G. Kaiser, "An Information Retrieval Approach For Automatically Construction Software Libraries," *IEEE Transaction on Software Engineering*, Vol. 17, No. 8, pp. 800-813, August 1991.

[3] 오효정, 맹성현, “점진적으로 계산되는 분류정보와 링크정보를 이용한 하이퍼텍스트 문서 분류 방법”, 정보과학회논문지 소프트웨어 및 응용 제29권 제7호, pp. 498-509, 2002.

[4] S. Park, and J. Palmer, "Automated Support to System Modeling from Informal Software Re-

quirements," *Proceedings of the 6th International conference on Software Engineering and Knowledge Engineering and Knowledge Engineering*, June 1994.
[5] 박수용, 서정연, 김학수, 고영중, "유사도 측정 기법을 이용한 효율적인 요구 분석 지원 시스템의 구현", 정보과학회 논문지 제27권 제1호, pp. 13-23, 2000.
[6] 강현규, "정보 검색", 정보처리 논문지 제5권 제5호, pp. 37-47, 1998.
[7] 이재윤, 최보영, 정영미, "문헌 자동분류에서 용어 가중치 기법에 대한 연구", 제7회 한국정보관리학회 학술대회 논문집, pp. 41-44, 2000.
[8] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval (Computer Series)*, New York: McGraw-Hill, 1983.
[9] 김명철, 김덕봉, 김유성, 김재훈, 박혁로, 이하규, 최신 정보검색론, 홍릉과학출판사, 2001.
[10] L. D. Baker and A. K. McCallum, "Distributional Clustering of Words for Text Classification," *Proceedings of SIGIR'98*, pp. 96-103, 1998.
[11] 정준호, 김미진, 이현주, 박미성, 이상조, "문서 요약 시스템을 위한 수사 구조 트리 생성", 한국정보과학회 가을 학술발표 논문집, Vol. 26. No. 2, pp. 175-177, 1999.
[12] M. Hajime, H. Takeo and O. Manabu, "Text Segmentation with Multiple Surface Linguistic Cues," *Proceedings of the COLING-ACL' 98*, pp. 881-885, August 1998.
[13] A. Jobbins and L. Evett, "Text Segmentation Using Reiteration and Collocation," *Proceedings of the COLING-ACL'98*, pp. 614-618, August 1998.
[14] H. Kozima, "Text Segmentation Based on Similarity between Words," *Proceedings of ACL'93*, pp. 286-288, January 1993.
[15] D. Litman and R. Passonneau, "Combining Multiple Knowledge Sources for Discourse Segmentation," *Proceedings of the 33rd ACL*, May 1995.
[16] A.I. Mel'cuk, *Dependency Syntax: Theory and Practice*, State Univ. of New York Press, 1988.
[17] 박수용, 서정연, 고영중, 강기선, 김재선, "요구사항 문장 범주화를 이용한 웹 기반의 요구 사항 추출 지원 시스템," 정보과학회 논문지 제27권 제4호, 2000.
[18] P. Hellwig, "Dependency Unification Grammar," *Proceedings of Colling86*, pp. 195-198, 1986.
[19] Y. Yaari, "Segmentation of Expository Texts by Hierarchical Agglomerative Clustering," *Proceedings of Ranlp'97*, pp. 135-142, September 1997.
[20] Y. Yang, "An evaluation of statistical approaches to text categorization," Information Retrieval Journal, May, 1999.
[21] 김명철, "공기기반 용어간 유사도를 이용한 정보검색 질의 확장 비교 연구", 박사논문, 한국과학기술원, 1999.
[22] Y. Karov and S. Edelman, "Similarity-based Word Sense Disambiguation," *Computational Linguistics*, Vol. 24, No. 1, pp. 41-60, March 1998.

김 혜 숙
1998년 2월 동신대학교 컴퓨터학과 졸업(학사). 2003년 8월 전남대학교 전자계산학과 졸업(이학석사). 2004년~현재 전남대학교 전자계산학과(박사과정). 관심분야는 패턴인식, 문서영상 정보검색

박 상 철
1999년 조선대학교 전자계산학과 졸업(학사). 2001년 조선대학교 대학원 전자계산학과 졸업(이학석사). 2003년~현재 전남대학교 대학원 전산학과(박사과정) 관심분야는 패턴인식, 영상처리, 컴퓨터 비젼

김 수 형
정보과학회논문지 : 소프트웨어 및 응용
제 31 권 제 9 호 참조