논문개요

잘 구축된 인터넷망과 컴퓨터기술 습득을 통해 누구나가 자신에게 유용한 정보를 찾아서 문제를 해결하거나 과제를 수행하는 것이 보편화 되었다. 그러나 수많은 정보를 제대로 활용하지 못하고 다른 사람의 자료를 전부 혹은 일부를 복사하거나 표절하는 경우가 많아지고 있다. 이에 대해 교수·평가자 입장에서 보자면 과제평가에 대한 어려움이 발생된다. 따라서 평가결과에 대한 신뢰도를 향상시키기 위해 다양한 색인추출방법과 유사도측정계산식에 따른 문서간 유사도 측정 시스템이 개발되고 있다.

본 논문은 기존의 색인어추출방법인 단어와 단어쌍 중에서 어떤 것이 정확도가 높으며, 또한 단어쌍 색인어추출방법에서는 추출된 단어쌍 개수가 정확도에 어떤 영향을 미치는지 알아보기 위해 새로운 색인추출방법을 제시하여 유사도 측정 계산식에 따른 정확도를 실험·분석하였다. 이 실험을 통해 정확도가 높은 색인추출방법을 찾아내어 교수·평가자의 평가결과에 대한 신뢰도를 높이고자 한다. 이를 위해 4가지 색인추출방법(단어색인, 명사그룹, 크기별 슬라이딩윈도우를 씌운 단어쌍(크기2-9), 가변적 슬라이딩 윈도우를 씌운 단어쌍)을 최소가중치계산식으로 유사도 측정 실험을 하였다.

실험 방법은 실제 교육대학원에서 과제로 내준 10페이지 분량의 독후감 10개를 1페이지로 분할하여 생성한 100개의 문서와 인터넷을 통해 검색한 독후감 400개 총 500개의 데이터를 대상으로 단어와 단어쌍 색인어추출방법 중에서 어느 것이 정확도가 더 높은지 알아보고, 단어쌍의 색인어의 개수가 정확도에 어떠한 영향을 미치는지에 대한 실험을 하였다. 그 결과 단어보다는 단어쌍 색인어추출방법이 높은 정확도를 보였으며, 단어쌍의 추출된 색인어의 개수가 정확도에 영향을 주어 정확도가 달라지는 것을 알 수 있었다.

목 차

표 목 차

그림 목차

Ⅰ. 서론

1. 연구배경 및 필요성

인터넷상의 정보와 다양한 서비스가 빠른 속도로 증가하고 있으며, 이를 만들고 사용하는 사람의 수 또한 증가하고 있다. 문서의 양과 사용자의 수가 증가함에 따라 방대한 양의 데이터를 분류하고 관리하는 일이 점점 중요한 도구가 되어가고 있다[1]. 웹의 활성화로 신속한 정보화가 진행 되면서 지식과 정보를 알고 이해하는 능력과 지식·정보 활용 능력이 개인의 능력을 가늠하는 척도가 되기도 한다. 이런 사회적 흐름과 잘 갖추어진 우리나라의 인터넷망을 이용해 학생들은 정보를 검색하고 주어진 과제를 해결 하고 있다. 인터넷망이 활성화 되어있지 않던 예전과 비교하면 많은 양의 정보를 활용해 수행한 과제이기에 질의 향상을 기대하지만, 실제로는 이를 역이용하여 과제 완성이나 문제 해결 과정에서 자신이 정보를 창출해 내는 것이 아니라 문서의 일부 혹은 전부를 다른 사람의 과제로 복사하거나 표절하는 경우가 많아졌다. 이러한 표절이나 과제 복사 행위들이 많아지면서 교수자의 과제평가 대한 어려움과 평가 결과에 대한 신뢰도가 떨어지는 현상이 일어나고 있다. 이에 문서 간 표절 여부를 판단 할 수 있는 문서비교시스템들이 개발되어 판단의 기준을 제시하고 있다.

본 논문은 이러한 문서비교시스템에서 기존에 사용되어진 색인추출방법 단어와 단어쌍 중에서 어떤 것이 정확도가 높으며, 단어쌍 색인추출 방법에서는 추출된 단어쌍의 개수가 정확도에 어떠한 영향을 미치는지를 비교·분석하는 실험을 통해 보다 효과적인 색인추출방법을 제시하고자 한다.

2. 연구내용 및 방법

본 논문은 기존에 사용되어진 색인추출방법 단어와 단어쌍 중에서 어떤 것이 정확도가 높으며, 단어쌍 색인추출 방법에서는 추출된 단어쌍의 개수가 정확도에 어떠한 영향을 미치는지를 비교·분석하여 보다 정확도가 높은 색인어 추출방법을 제시하여 교수·평가자의 평가결과에 대한 신뢰도를 높이고자 한다. 이를 위해 실험에 사용된 4가지 색인추출방법을 최소가중치 유사도 측정 계산식을 사용하여 실험을 하였다.

색인어 추출을 위해 형태소 분석기를 사용하였으며, 형태소 분석기를 통해 얻어진 명사만을 대상으로 색인어를 구성하였다. 실험에 사용한 색인어는 크게 단어와 단어 쌍으로 구분되며, 단어 쌍은 다시 3가지 방법을 이용하여 구성하였다.

문서간 유사도 정확성의 측정을 위해 실제 교육대학원에서 과제로 내준 10페이지 분량의 독후감 10개를 1페이지로 분할하여 생성한 100개의 문서와 인터넷을 통해 검색한 독후감 400개 총 500개의 데이터를 대상으로 크게 2그룹으로 나누어 실험을 하였다. 첫 번째 그룹은 500개의 문서를 10:1부터 10:10까지 분할하여 유사도를 측정하여 나온 결과의 정확성을 알아보기 위해 실험을 하였으며, 두 번째 그룹은 500개의 문서를 각각 같은 주제의 다른 문서와 9:1, 5:5 의 비율로 섞은 것과 각각의 문서의 내용에 변화를 주지 않고 문단의 순서를 마구잡이로 바꾸어 측정한 유사도 결과가 실제로 변화를 준 것처럼 나오는지 여부를 판단하는 실험을 하였다.

본 논문의 2장에서는 문서간 유사도 측정과 관련된 기존 연구에 대해 고찰하고, 3장에서는 형태소분석기를 통해 얻어진 명사를 이용하여 구성한 색인어의 추출방법을 설명하고, 이에 따른 유사도 시스템의 알고리즘을 설계한다. 4장에서는 문서간 유사도 측정에 따른 실험·분석 결과를 기술한다. 마지막으로 5장에서는 연구결과의 요약과 기대하는 효과, 앞으로 보완해 나가야 할 점을 제안한다.

Ⅱ. 관련연구

본 장에서는 본 연구의 배경이 되는 색인어 추출과 단어 빈도 가중치, 유사도 측정계산식에 대해 살펴본다.

1. 색인어 추출

색인은 특정한 정보가 필요한 사람에게 그 정보의 위치를 지시해 주는 역할과 방대한 정보원으로부터 가장 유사한 내용의 정보 자료만을 선별해 주는 역할을 한다. 또한 문서의 내용을 나타내거나, 그 문서를 다른 문서들로부터 구별할 수 있도록 그 문서의 선택 단서가 되는 단어 또는 단어구를 의미한다. 문장의 내용이나 특성을 잘 반영하는 명사, 동사, 형용사 등을 내용어(content word)라고 하며 이런 내용어가 색인어가 된다[6].

색인어 추출 방법은 형태에 따라 한 단어가 하나의 색인을 구성하는 방법과 여러 단어로 이루어진 하나의 구가 색인이 되는 방법으로 구분된다. 색인어가 단일 단어로 구성한 것은 적은 양의 데이터에서도 많은 색인을 추출할 수 있다는 장점이 있는 반면에 문맥정보를 포함할 수 없다는 단점이 있다. 이에 반해 단어쌍으로 구성된 색인은 공기정보 등을 이용해서 단어의 쌍을 색인으로 보는 것으로 문맥 정보를 어느 정도 반영할 수 있다는 장점이 있다[12,13].

본 논문에서는 단어와 단어쌍 색인추출방법에 따른 문서간 유사도 측정 결과의 정확도를 비교・분석하고자 한다. 따라서 단어 색인 생성 방법과 단어쌍 색인 생성 방법에 대해 연구하고자 한다.

1) 단어 색인 생성

일반적으로 유사한 단어는 유사한 문맥(content)에 위치하는 경향이 있으므로 단어의 이런 속성에 의한 문장간 유사도 측정을 통해 이를 문서로 확장한다. 단어와 문장은 상호 보충적인 역할을 수행한다. [2]에 의하면 특정한 단어가 문서 집단속의 상호관련 없는 문서들을 분리시키는 능력치가 큰 것이 좋은 색인어가 되고 나쁜 색인어일수록 상호 관련이 없는 문서들을 묶어준다고 하였다. 단어색인 방법에서는 단순히 형태소 분석기를 통해 추출된 명사가 색인어가 된다. 그림 2.1은 형태소 분석기를 통해 얻어진 결과이다. 그림 2.1에서 단어 뒤에 /(NN)는 명사를 뜻한다.

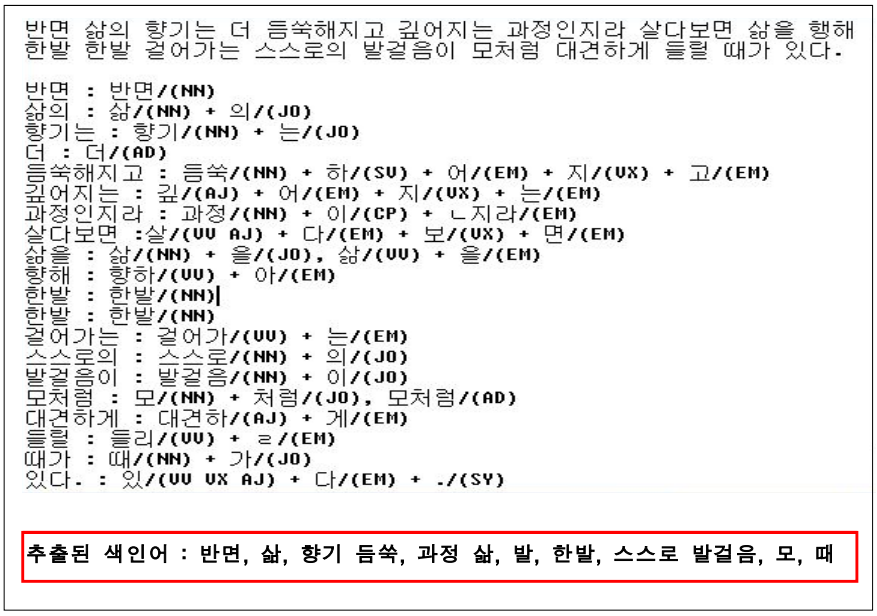


그림 2.1 형태소 분석기를 통해 추출된 명사 색인어

2) 단어쌍 색인 생성

색인 추출에 문맥 정보를 반영하기 위한 방법으로 인접한 단어 사이의 공기 정보가 그동안 많이 사용되어 왔다. 인접한 단어 사이의 공기 정보를 추출하기 위해서 슬라이딩 윈도우 기법[9]를 사용하는데, 다음과 같은 단계를 거쳐 수행된다. 먼저 문장을 형태소분석기를 통해 추출된 내용어(명사, 동사, 형용사 등)를 순서 열에 일정 크기의 윈도우(window)를 설정하고, 맨 앞의 내용어와 다음 내용어들간의 쌍을 추출한다. 윈도우는 문장의 처음부터 마지막 내용어까지 움직이며, 크기는 문장의 끝에서 문장의 경계를 넘지 않도록 줄어든다. 문장의 끝에서 윈도우의 크기를 줄이는 이유는 다른 문장에 속해 있는 내용어가 같은 문자내의 내용어 보다 약한 문맥정보를 가지기 때문이다. 그리고 윈도우의 슬라이딩(sliding)을 한 문장으로 제한하여 추출되는 색인의 수를 적절한 수준으로 유지하기 위함이다[6].

그림 2.2는 슬라이딩 윈도우 기법을 이용하여 내용어중 명사만을 대상으로 색인어를 추출한 예이다.

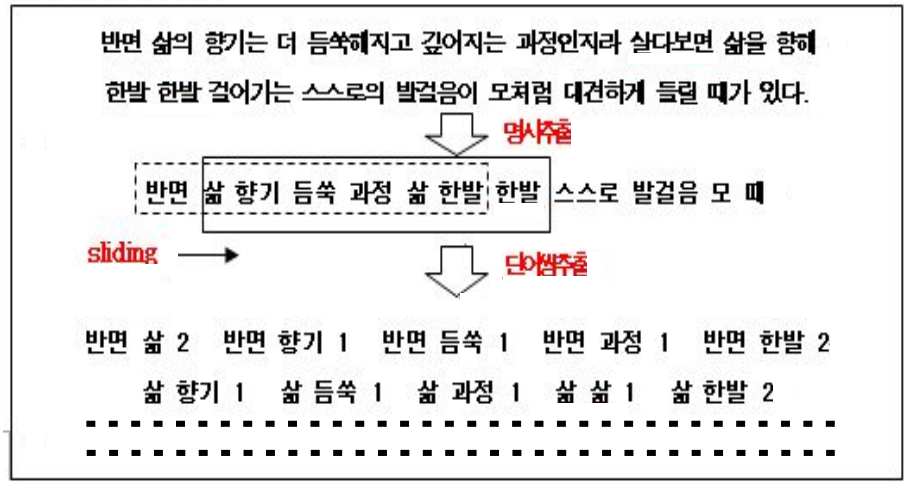


그림 2.2 슬라이딩 윈도우 기법을 이용한 명사 색인어 추출의 예

2. 단어 빈도 가중치

단어빈도(Term Frequency)란 단어가 문서 내에 몇 번 출현했는지를 나타내는 빈도수를 말하며, 단어빈도가 높을수록 그 단어가 문헌의 주제를 대표할 확률이 높다는 통계적 기법의 기본적인 가설에 근거하여 단어빈도만을 이용하여 가중치(주제어로서의 중요도)를 계산한다[2]. 다음은 다양한 단어 빈도 가중치 공식을 보여주고 있다[3]. tf는 한 단어가 문서 내에서 나타난 단어빈도수이며, TF는 tf에 따른 가중치 값이다.

1) 이진 TF : 단어가 출현한 경우를 모두 1로 지정하여 가중치를 주는 공식.

$$TF = 1\ (\mathrm{if} > 0)\ ,\ 0\ (otherwise)$$

2) 단순 TF : 단어빈도 tf가 가중치를 나타낸다.

$$TF = tf$$

3) 로그 TF : tf가 1인 단어의 지나치게 낮은 영향력을 보충, tf가 높은 단어의 지나친 영향력을 낮추기 위해 TREC-1에서 SMART팀이 제안한 공식.

$$TF = 1 + \log(tf)$$

4) 더블로그2 TF : tf가 0,1,2일 때는 가중치도 0,1,2가 되고 tf가 3이상일 때는 가중치가 로그곡선을 따르도록 하는 공식.

$$TF = 1 + \log_2(1 + \log_2(tf))$$

4) 루트 TF : 로그 TF와 같은 효과를 가지지만 tf가 높은 경우의 의미를 로그 TF보다는 덜 축소하는 공식

$$TF = \sqrt{tf}$$

5) 보정 TF : 가중치를 일정 범위로 한정시켜서 최소 빈도의 단어라도 일정 값 이상이 되도록 하면서 동시에 최대값도 제한하는 공식.(w는 일정 값 이상이 되도록 하는 보정 값을 나타낸다.)

$$TF = (1-w) + w \times \frac{tf}{\max tf}$$

6) Okapi TF : 2-포아송 모형을 적용하는 Okapi시스템에서 사용하는 공식

$$TF = \frac{tf}{2 + tf}$$

7) 루트직선 TF : 루트 TF공식과 가중치 선의 기울기가 유사하도록 tf가 1일 때와 9일 때의 값이 루트 TF와 같은 직선 공식.

$$TF = \frac{tf+3}{4}$$

8) 더블로그 TF : 질의가 소소의 질의어로 구성된 경우에는 로그 TF로도 tf가 높은 단어의 지나친 영향력을 낮추는 것이 불충분 하다고 판단하여 Singhal, et al.(1998)은 TREC-7에서 로그를 두 번 취하는 공식을 제안.

$$TF = 1+\log(1+\log(tf))$$

이상의 여러 가지 단어빈도 가중치 공식을 그래프로 나타내면 그림 2.1과 같다.

가중치 선의 기울기는 이진 TF가 0으로 가장 낮고 단순 TF가 1로 가장 높다[4]. 단순TF가 저빈도어에 대한 가중치보다는 고빈도어에 대한 가중치 값이 상대적으로 높게 나타나 결국 고빈도에 대한 식별력이 우수하게 나타나는 것을 알 수 있다[5].

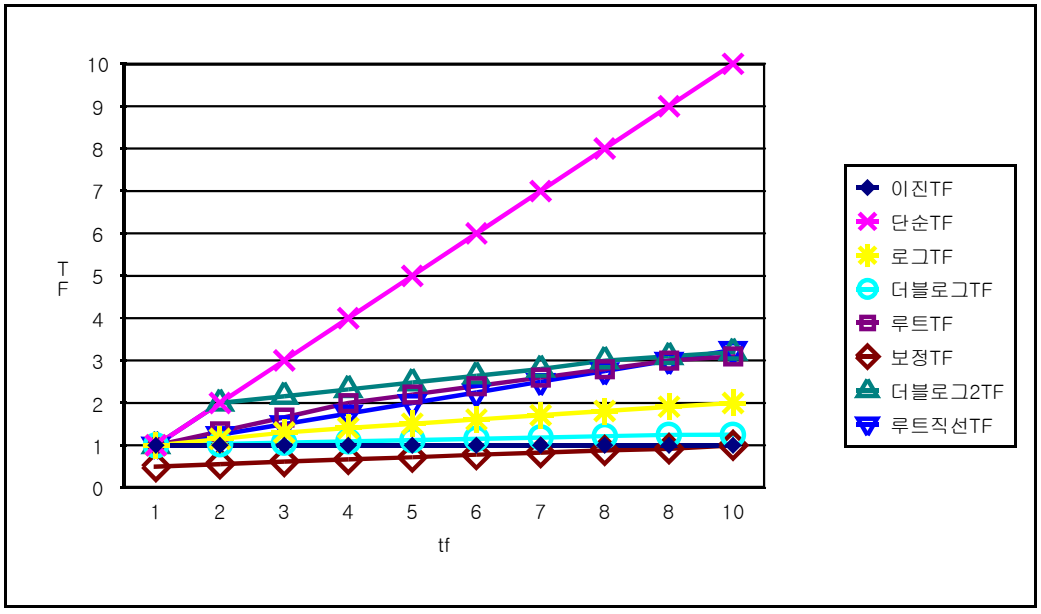


그림 2.3 tf에 따른 각 공식 값의 변화

3. 유사도 측정 계산식

문서간 혹은 문서와 클러스터간 유사도 계산에 사용되는 계산식으로 여러 가지 계산식(Cosine 가중치, 최소 가중치, 비율 가중치, 곱 가중치, log 가중치 중 최소값, 최소가중치에 Log취한 값)이 있으며, 간소화와 정규화 기능을 갖고 있어서 대부분의 문서 간 유사도 계산에 사용된다[4]. 다음은 다양한 유사도 측정 계산식을 보여주고 있다. tf_j는 문서 j에서 추출된 전체 색인어에 대한 빈도수를 뜻하며, tf_k는 문서 k에서 추출된 전체 색인어에 대한 빈도수를 뜻한다. tf_{ij}와 tf_{ik}는 각각 문서 j와 k의 공통색인어 i에 대한 빈도수를 뜻한다.

1) Cosine 가중치 : 공통된 색인의 수가 같을 때 각 문서에서 색인의 수가 적을수록 높은 유사도를 갖는 것으로 긴 문서의 경우 색인의 수도 많으며 당연히 공통 색인수도 많아지기에 이 경우 불이익을 주기 위한 공식.

$$\frac{\sum_i (tf_{ij} \times tf_{ik})}{\sqrt{\sum_j tf_j^2} \cdot \sqrt{\sum_k tf_k^2}}$$

2) 최소 가중치 : 두 문서에서 공통되는 각 단어에 대한 빈도수를 각 문서의 전체 단어에 대한 빈도수의 합으로 나눈 값 중 최소치를 취한 다음, 이를 모두 합하는 공식.

$$\sum_i \min\left(\frac{tf_{ij}}{\sum_j tf_j}, \frac{tf_{ik}}{\sum_k tf_k}\right)$$

3) 비율 가중치 : 두 문서에서 공통되는 단어에 대한 빈도수를 각 문서의 빈도수 중 최대치로 나눈 값을 취한 다음 이를 모두 합하는 공식으로 자주 등장하는 단어는 문서의 유사도나 의미의 정보량을 적게 지니게 하는 속성이 있다.

$$\sum_{i}(\min(\frac{tf_{ij}}{\max tf_{j}},\frac{tf_{ik}}{\max tf_{k}})$$

4) 곱 가중치 : 각 단어에 대한 가중치의 곱을 취하는 계산식으로 정보검색에서 가장 빈번하게 사용하는 계산식 중의 하나이다.

$$\sum_{i}(\frac{tf_{ij}}{\sum_{j}tf_{j}}\times\frac{tf_{ik}}{\sum_{k}tf_{k}})$$

5) Log가중치중 최소값 : 한 단어에 대해 지나치게 큰 가중치의 영향을 최소화 시키고자 일치하는 각 단어에 대한 가중치에 로그를 취한 공식.

$$\sum_{i}\min(\log\frac{tf_{ij}}{\sum_{j}tf_{j}},\log\frac{tf_{ik}}{\sum_{k}tf_{k}})$$

6) 최소가중치에 Log취한 값 : 최소가중치 계산식에서 구해진 최소가중치의 지나치게 큰 영향을 최소화 시키고자 가중치에 로그를 취한 공식

$$\log\left[\sum_{i}\min(\frac{tf_{ij}}{\sum_{j}tf_{j}},\frac{tf_{ik}}{\sum_{k}tf_{k}})\right]$$

Ⅲ. 유사도측정 시스템 설계

본 장에서는 색인추출 방법에 따른 문서간 유사도측정 시스템의 전체적인 구조와 4가지 색인추출과정에 따라 유사도 측정 정확도가 어떻게 달라지는지에 대해 기술하고자 한다.

본 논문의 실험에서 사용된 색인은 크게 단어와 단어쌍으로 구성되어 있다. 단어쌍 색인은 다시 3가지 색인추출방법(명사그룹, 가변적 윈도우 슬라이딩, 크기별 윈도우 슬라이딩)으로 생성하여 총 4가지 형태의 색인을 추출하여 실험 하였다. 전체적인 시스템 구조도는 그림 3.1과 같다.

1. 전체 시스템 구조도

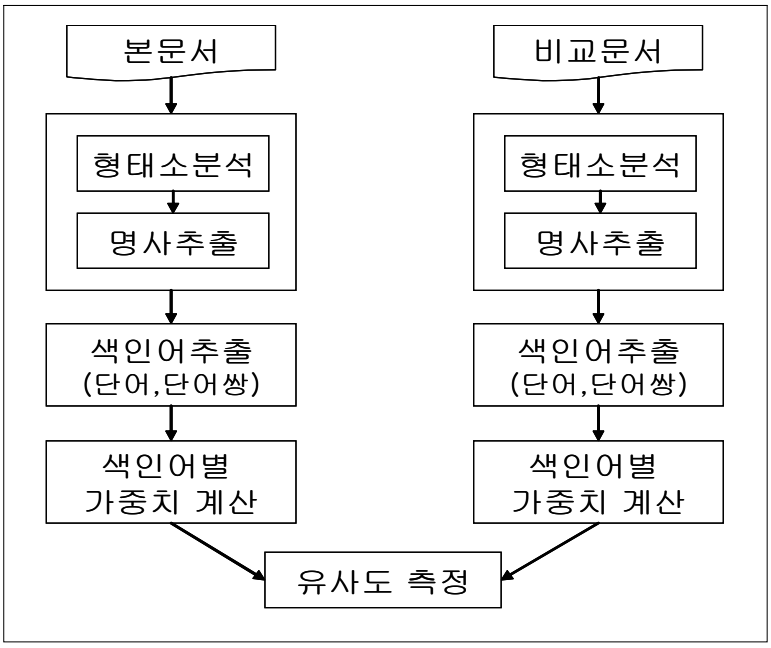


그림 3.1 전체 시스템 구조도

본 시스템에서는 본문서와 비교문서에서 형태소분석기를 통해 명사만을 추출하여, 추출한 명사로 각기 다른 색인어추출 방법에 따라 색인어를 생성한다. 생성된 색인어는 색인어별 가중치를 계산하고 이를 이용하여 유사도 측정계산식에 따른 문서간 유사도를 측정하여 기존의 색인추출방법인 단어와 단어쌍 중에서 어느 것이 정확도가 더 높은지를 알아보았다. 또한 단어쌍 색인추출방법에서 추출된 색인어의 개수가 정확도에 어떤 영향을 미치는지 알아보기 위해 2가지 색인추출방법을 제안하여 실험·분석하였다. 유사도 측정 계산식으로는 최소가중치 계산식을 이용하였다.

2. 색인어추출

본 논문에서는 형태소 분석기[7]을 통해 문서 내에서 내용어로서의 가치를 가진 명사들을 추출하여 이를 토대로 색인어를 생성하였다. 색인어 추출은 총 4 가지 방법을 이용하였으며, 첫 번째는 기존에 많이 사용하고 있는 하나의 단어만을 색인어로 보는 단어 색인어 추출 방법을 적용하였고, 두 번째는 기존에 단어쌍 색인어 추출 방법에서 많이 사용 되어진 슬라이딩 윈도우를 적용하여 윈도우의 크기를 2에서 9까지 변화를 주어 색인어를 추출하는 방법을 적용하였다. 세 번째 방법은 문장단위로 추출된 모든 명사들을 하나의 단어쌍 색인어로 보는 방법이고, 네 번째 방법은 한 문장 내에서 가능한 모든 쌍들을 추출하기 위해 슬라이딩 윈도우를 씌워 색인어를 추출하는 방법이다. 세 번째와 네 번째 방법은 본 논문에서 제시하는 방법이다. 그림 3.2는 실제 형태소 분석기를 통해 얻어진 결과이다.

```
반면 삶의 향기는 더 흠쑥해지고 쿶어지는 과정인지라 살다보면 삶을 행해
한발 한발 걸어가는 스스로의 발걸음이 모처럼 대견하게 들릴 때가 있다.

반면 : 반면/(NN)
삶의 : 삶/(NN) + 의/(JO)
향기는 : 향기/(NN) + 는/(JO)
더 : 더/(AD)
흠쑥해지고 : 흠쑥/(NN) + 하/(SV) + 어/(EM) + 지/(VX) + 고/(EM)
쿶어지는 : 쿶/(AJ) + 어/(EM) + 지/(VX) + 는/(EM)
과정인지라 : 과정/(NN) + 이/(CP) + ㄴ지라/(EM)
살다보면 :살/(VV AJ) + 다/(EM) + 보/(VX) + 면/(EM)
삶을 : 삶/(NN) + 을/(JO), 삶/(VV) + 을/(EM)
행해 : 행하/(VV) + 아/(EM)
한발 : 한발/(NN)
한발 : 한발/(NN)
걸어가는 : 걸어가/(VV) + 는/(EM)
스스로의 : 스스로/(NN) + 의/(JO)
발걸음이 : 발걸음/(NN) + 이/(JO)
모처럼 : 모/(NN) + 처럼/(JO), 모처럼/(AD)
대견하게 : 대견하/(AJ) + 게/(EM)
들릴 : 들리/(VV) + ㄹ/(EM)
때가 : 때/(NN) + 가/(JO)
있다. : 있/(VV VX AJ) + 다/(EM) + ./(SY)
```

그림 3.2 형태소분석기를 통해 얻어진 결과

1) 단어 색인 생성

기존의 하나의 단어를 색인어로 생성할 때 많이 사용하는 방법으로 형태소분석기를 통해 얻어진 각각의 명사들 하나가 색인어가 되도록 하였다. 그림 3.3은 실제로 단어 색인 생성 방법으로 색인어를 추출하여 저장한 결과화면이다.

편집	ID	WORD	W_COUNT	W_ROW	W_COL	W_TYPE
	20041	선생님	-	0	-	NN
	20041	지금	-	1	-	NN
	20041	내	-	2	-	NN
	20041	안	-	3	-	NN
	20041	사람	-	4	-	NN
	20041	대학교	-	5	-	NN
	20041	진학	-	6	-	NN
	20041	이전	-	7	-	NN
	20041	학창시절	-	8	-	NN
	20041	생활기록부	-	9	-	NN
	20041	“장래희망	-	10	-	NN
	20041	직업	-	11	-	NN

그림 3.3 단어 색인 생성 방법을 통해 추출된 색인

2) 단어쌍 색인 생성

하나의 단어만을 대상으로 색인어를 생성하는 방법은 문서의 문맥 정보를 반영할 수 없다. 따라서 문서의 문맥 정보를 반영하기 위해 색인어를 단어쌍으로 생성한다. 먼저 형태소 분석기[7]를 통해 얻어진 명사들의 순서열에 일정한 크기의 윈도우를 설정하고, 맨 앞의 명사와 다음 명사들 간의 쌍을 추출한다. 윈도우는 문장의 처음에서부터 마지막 내용어까지 움직이며, 크기는 문장의 끝에서 문장의 경계를 넘지 않도록 줄어든다.

그림 3.4는 실제로 단어쌍 색인 생성 방법으로 색인어를 추출하여 저장한 결과화면이다.

편집	DOC_ID	WORD_ONE	WORD_TWO	WORD_THREE
	10581	인생	수	인생수
	10581	도움	수	도움수
	10581	아이	능력	아이능력
	10581	아이	칭찬	아이칭찬
	10581	아이	칭찬	아이칭찬
	10581	능력	칭찬	능력칭찬
	10581	능력	칭찬	능력칭찬
	10581	능력	아이	능력아이
	10581	칭찬	칭찬	칭찬칭찬
	10581	칭찬	아이	칭찬아이
	10581	칭찬	영향	칭찬영향
	10581	칭찬	아이	칭찬아이
	10581	칭찬	영향	칭찬영향
	10581	칭찬	줄	칭찬줄
	10581	아이	영향	아이영향

1-15/500행 초과

그림 3.4 단어쌍 색인 생성 방법을 통해 추출된 색인어

① 슬라이딩 윈도우를 이용한 색인 생성

기존의 단어쌍 색인 생성시 많이 사용하는 방법 중에 하나가 슬라이딩 윈도우를 씌워 단어쌍을 추출하는 것이다. 따라서 본 논문에서도 슬라이딩 윈도우를 적용하여 단어쌍 색인어를 추출하였는데, 윈도우의 크기에 따라 정확도에 어떤 영향을 미치는지 알아보기 위해 형태소 분석기를 통해 얻어진 문장단위의 명사들에 슬라이딩 윈도우의 크기를 2에서 9까지 변화를 주어 단어쌍 색인어를 추출하였다. 그림 3.5과 그림 3.6은 슬라이딩 윈도우의 크기를 각각 3과 4로 하였을 때 생성되는 색인어의 예이다.

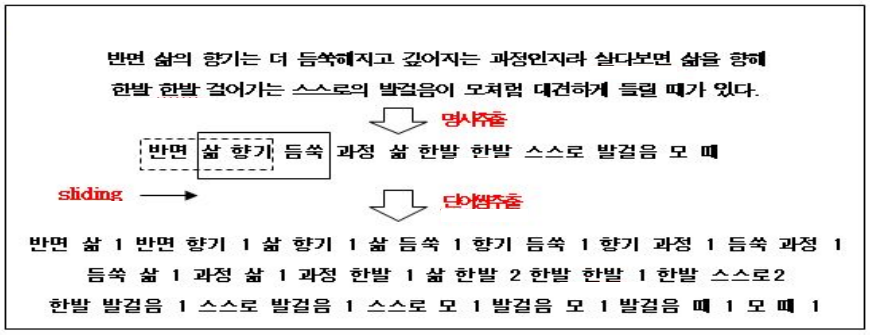


그림 3.5 윈도우 크기 3을 적용하여 추출한 단어쌍 색인어의 예

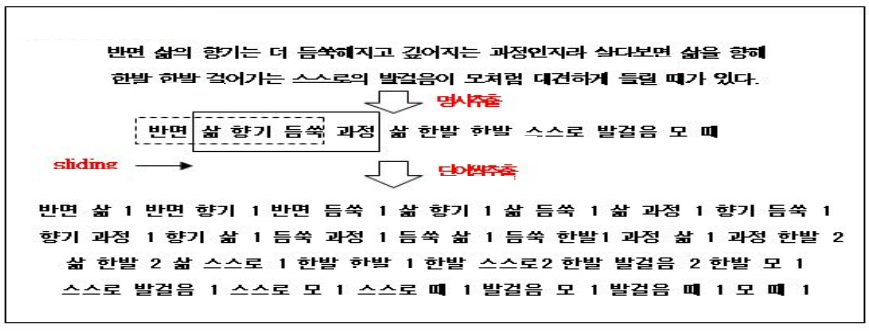


그림 3.6 윈도우 크기 4를 적용하여 추출한 단어쌍 색인어의 예

② 문장 내 추출 가능한 모든 단어쌍 색인 생성

문장 내 추출 가능한 모든 단어쌍 색인 생성 방법은 본 논문에서 제안하는 방법으로 슬라이딩 윈도우의 크기를 고정적인 크기를 주는 것이 아니라 문장단위로 추출된 명사들 전체에 윈도우를 씌워 단어쌍을 추출한 후 맨 처음에 있는 명사 하나를 제외한 명사들에 다시 윈도우를 씌워 색인어를 추출하는 것이다. 즉 문장의 끝까지 윈도우 크기를 하나씩 줄여 색인어를 생성하는 방법이다.

이는 한 문장 내에서 추출 가능한 모든 단어쌍을 추출 하는 방법으로 문장 단위로 추출된 색인어의 수가 유사도 측정 정확도에 어떤 영향을 미치는지 알아보기 위해 제안된 방법이다. 그림 3.8은 실제 가변적 슬라이딩 윈도우를 이용해 얻어진 색인어의 예이다.

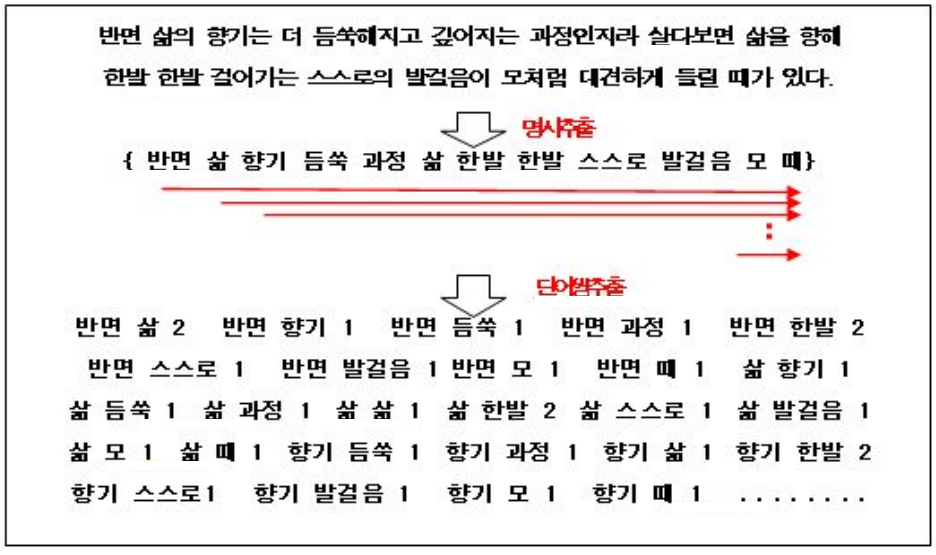


그림 3.8 가변적 슬라이딩 윈도우를 이용한 색인 생성방법의 예

3) 단어 그룹 색인 생성

단어 그룹 색인 생성 방법은 본 논문에서 제안하는 방법으로 형태소 분석기를 통해 문장 단위로 추출된 모든 명사들을 하나로 묶은 명사 그룹이 하나의 색인어가 되게 하는 것이다. 이 방법은 인터넷 등을 통해 검색된 자료를 재해석 하거나 변경하지 않고 과제 수행자(학생)가 기존의 문서를 그대로 사용하여 과제를 작성한 경우의 문서간 유사도 측정을 위해 제안된 방법이다. 실제 주변의 사례를 살펴보면 과제 수행자가 과제를 작성할 때 검색한 자료를 변경하지 않고 그대로 가져오거나 인터넷 검색을 통해 기존에 작성된 같은 주제의 다른 사람의 과제를 일부 혹은 전부 복사하여 작성하는 경우가 많다. 따라서 이와 같은 색인 생성 방법을 제안하게 되었다. 그림 3.7은 실제 단어 그룹 색인 생성 방법을 통해 얻어진 색인어의 예이다.

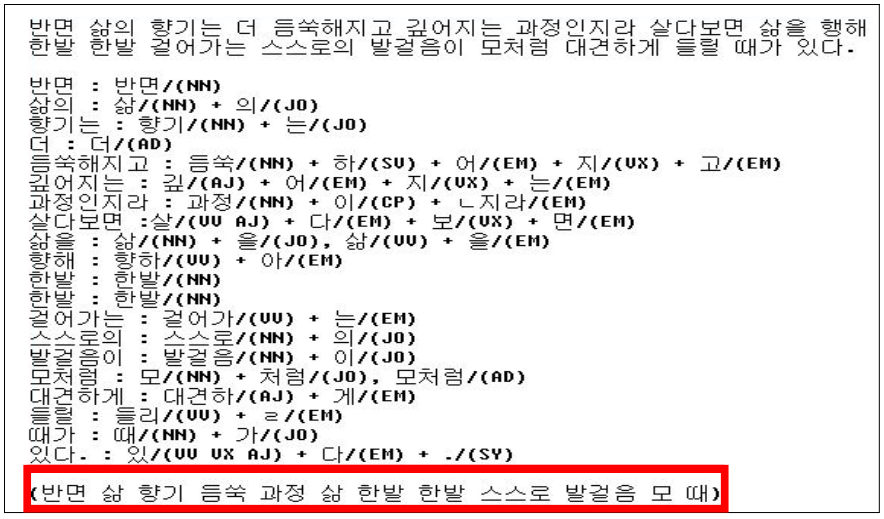


그림 3.7 단어 그룹 색인 생성의 예

3. 유사도 측정

본 시스템에서는 유사도 측정을 위해 색인어별 가중치 적용하였는데, 용어 가중치와 관련된 선행연구들을 살펴본 결과 단어빈도 가중치 계산식으로 단순 TF보다 기울기가 낮은 로그 TF나 루트 TF가 좋은 것으로 나타났으나, [5]가 이를 검토하여 실험 분석한 결과 단순 TF의 성능이 다른 가중치 계산식과 별 차이가 없는 것으로 나타났다. 이에 본 논문에서는 다른 가중치 계산식에 비해 간단하지만 성능에는 별 차이가 없는 단순 TF를 사용하여 시스템을 설계하였다.

유사도 측정 계산식으로 두 문서에서 공통되는 단어에 대한 빈도수를 각 문서의 전체 단어에 대한 빈도수의 합으로 나눈 값 중 최소치를 취한 다음에 이를 두 문서에 대해 공통으로 존재하는 단어에 대한 가중치를 모두 합해서 결과 값을 얻는 최소 가중치 계산식을 사용하였다. 이는 최소 가중치 계산식이 [2]에서 7가지 유사도 측정 계산식에 대한 비교・분석을 한 결과 가장 높은 정확도를 보였기 때문이다. tf_{ij} 와 tf_{ik}는 각각 j와 k문서에 공통으로 존재하는 공통색인어 i의 빈도수를 뜻하며, $\sum tf_j$와 $\sum tf_k$는 j와 k문서의 전체 색인어 빈도수의 합을 나타낸다. $\sum$min은 j와 k문서의 공통 색인어들을 j와 k문서의 전체빈도수로 나누어 나온 값 중에서 적은 것들을 합하라는 것을 의미한다.

$$\sum_i \min \left(\frac{tf_{ij}}{\sum_j tf_j} , \frac{tf_{ik}}{\sum_k tf_k} \right)$$

Ⅳ. 실험 및 결과 분석

본 장에서는 4가지 색인어 추출 방법에 따른 문서간 유사도 측정 시스템의 실험하여 정확률을 비교·분석한 결과에 대해 기술한다.

본 논문에서 실험에 사용한 색인어 추출방법은 총 4가지로 단어 색인어 생성 방법, 슬라이딩 윈도우를 이용한 색인어 생성 방법, 문장 내 추출 가능한 모든 단어쌍 색인어 생성 방법, 단어그룹 색인어 생성방법이다.

1. 개발환경

본 논문에서 설계한 문서간 유사도 측정 시스템의 구현에 사용되는 개발환경은 다음 표 4.1과 같다.

구분		사양
Software	OS	Microsoft Windows XP Processional
	DBMS	Oracle Database 10g Express Edition
	Language	Java
	Language Editor	Eclipse SDK
Hardware	CPU	Intel Core 2 Duo E6420, 2133 MHz
	RAM	1GB
	HDD	150GB

표 4.1 시스템 개발환경

2. 실험문서 집단 구성 방법

실험문서는 크게 두 가지 방법에 의해 구성하였는데, 첫 번째 실험문서 집단은 10페이지 분량의 독후감 10개를 1페이지로 분할하여 생성한 100개의 문서와 인터넷을 통해 검색한 독후감 400개의 문서를 각각 1:9, 2:8, 3:7 4:6, 5:5, 6:4, 7:3, 8:2, 9:1로 분할하여 실험문서를 구성하였다. 이는 4 가지 색인어 추출 방법에 따른 정확도를 비교하기 위해 구성된 실험문서 집단으로 문서의 분할된 정도의 정확한 기준은 단어 그룹 색인 생성방법에 의해 측정된 결과로 하였다. 이는 비교문서의 일부를 변화 시킨 것이 아니라 단지 문장 단위로 분할하였기에 단어그룹 색인 생성 방법에 의해 측정된 유사도를 곧 분할된 비율로 보았기 때문이다. 그림 4.1은 첫 번째 실험문서 집단의 분할된 형태를 그림으로 나타낸 것이고, 그림 4.2는 첫 번째 실험문서 집단의 문서 비교 방법을 나타낸 것이다.

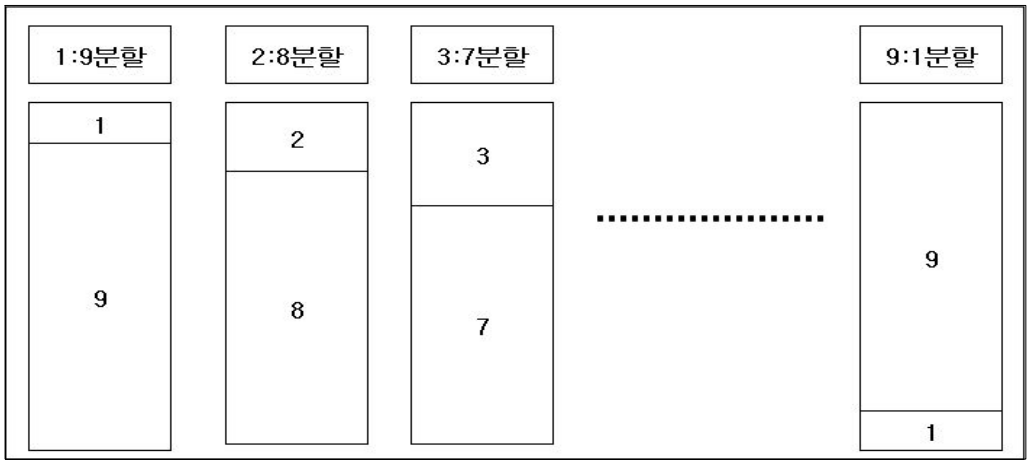


그림 4.1 첫 번째 실험문서 집단의 문서 분할 형태

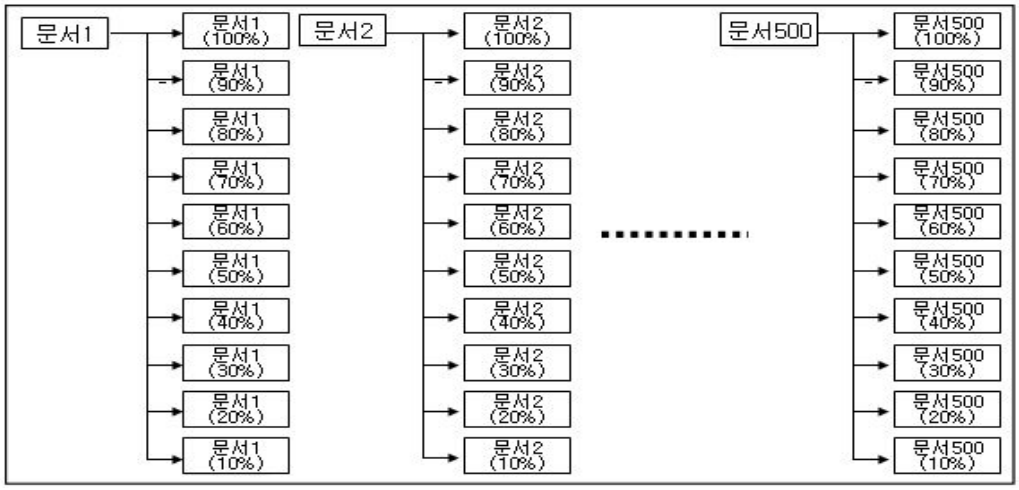


그림 4.2 첫 번째 실험문서 집단의 문서 비교 방법

- 『내 안의 빛나는 1%를 믿어준 사람』 을 읽고
1. 내 안에 깃들인 선생님의 기억을 찾아서
사람들 모두가 이런 잊을 수 없는 선생님 한 두 분 정도를 떠올려 볼 수 있을 것이다. 그들은 우리가 힘겨울 때 언제나 힘이 되어
교사의 역할 중 또한 중요한 것이 있다면 그것은 바로 아이들의 잠재력을 밖으로 이끌어 내어 주는 것이다. 교사는 말 한마디, 행동
이 책을 읽는 동안에 이런 생각들과 함께 나의 학창시절이 떠올랐고 나에게도 믿음을 주고, 용기를 주며 나의 잠재력을 밖으로 끌어
2. 나의 빛이 되어주신 선생님
어렸을 때, 자세히 말하자면 내가 처음 중학교에 입학했을 때 나의 담임선생님께서는 나조차도 잘 모르고 있던 나에 대해서 새로이
움은 쉬이 사그라들지 않았다.
그러던 나에게, 담임선생님께서는 한줄기 빛처럼 다가오셨다. 수학을 가르치셨던 담임선생님께서는 다정다감한 성격은 아니셨지만
그 이후로도 이런 빛과 같은 선생님을 한 분 더 만난 적이 있다. 그때는 고등학교에 갓 입학하였을 때였는데, 외국어고등학교에 진
3. 아이들의 빛이 되어주기 위해서
미래 초등교사의 한 사람으로서 이 책을 읽는 내내 선생님의 역할에 대해서 다시 한 번 진지하게 생각해 보게 되었다. 내가 이제까
첫 번째로, 교사는 지식 전달의 역할을 제대로 해야 하며 아이들이 학과목에 흥미를 갖도록 해야 한다. 정확한 정보를 전달하는 것
여기서 나의 경험 또한 빼놓을 수가 없는데, 나는 원래 과학이란 과목을 좋아하지 않았지만 선생님의 영향으로 과학을 좋아하게 되
두 번째로, 교사는 상담가의 역할을 해야 한다. 힘들고 어려운 처지에 처해있는 아이들의 어려움을 미리 파악할 수 있어야 하며,
또 하나, 상담가로서의 교사에게서는 학생이 먼저 상담을 요청하기 전에 그 학생이 현재 문제를 가지고 있는지 꿰뚫어 보는 통찰력
세 번째, 교사는 자신을 모두 내 보여 줄 수 있는 사람이어야 한다. 교사가 학생들 앞에서 자신의 솔직한 모습을 보이지 않는다면
서로의 마음과 마음은 통한다. 교사가 자신의 마음을 모두 보여주지 않고 자신의 마음의 문을 닫아버린다면 아이들도 그것을 느낄
넷째로, 교사는 개개인의 능력을 믿고, 이것을 일깨워 줄 수 있어야 한다. 아이들은 자신들이 가진 잠재된 능력을 깨닫지 못하는 경
믿음 이라는 것은 어느 순간에나 큰 힘을 발휘하는 것 같다. 나를 아무도 믿어주는 사람이 없다고 생각하는 것과 누군가가 나를 보
앤드류 카네기의 '교육의 가장 위대한 성과는 아이들로 하여금 스스로 배우도록 도와주는 것이다' 라는 말과 현대의 서양 교육의 목
다섯째, 교사는 학생들에게 깊은 관심을 보여주는 사람이 되어야 하며, 모두가 자신이 특별한 존재라고 느끼도록 해 주어야 한다.
그렇다면 문제아라고 여겨지는 아이에게 가져야 하는 관심은 어떤 종류의 관심일까? 이 문제의 해답은 간단하다. 그저 다른 아이들
아이들은 모두 자기가 특별한 존재이기를 바란다. 그리고 교사는 그들을 그렇게 만들어 줄 수 있다. 모두는 특별한 존재이며 자신이
4. 모두의 1%를 믿는 선생님 되기
교사의 이러한 역할들을 생각해 보면서 내가 후에 교사가 된다면 이런 역할들을 훌륭히 소화해 내는 교사가 될 수 있을까 염려스러
단순히 학과 공부와 바로 내 앞의 미래만을 좇던 나를 이 책이 변화시킨 것이다. 교육은 한 사람을 그 전의 모습으로부터 '변화'

그림 4.3 첫 번째 실험문서 집단의 10:10의 실제 예

- 『내 안의 빛나는 1%를 믿어준 사람』 을 읽고
1. 내 안에 깃들인 선생님의 기억을 찾아서
사람들 모두가 이런 잊을 수 없는 선생님 한 두 분 정도를 떠올려 볼 수 있을 것이다. 그들은 우리가 힘겨울 때 언제나 힘이 되어
교사의 역할 중 또한 중요한 것이 있다면 그것은 바로 아이들의 잠재력을 밖으로 이끌어 내어 주는 것이다. 교사는 말 한마디, 행동

그림 4.4 첫 번째 실험문서 집단의 10:1의 실제 예

두 번째 실험문서의 구성은 수집한 독후감 10페이지 분량의 10개를 한 페이지 단위로 나눈 총 100개의 문서와 인터넷을 통해 검색한 독후감 400개를 각각 본문서 90%와 수집한 독후감 중 다른 문서 10%를 결합한 형태와 본문서 50%와 수집한 독후감 중 다른 문서 50%를 결합한 형태, 그리고 각 문서 내에서 단지 문단의 순서를 뒤바꾼 형태의 문서로 구성하였다.

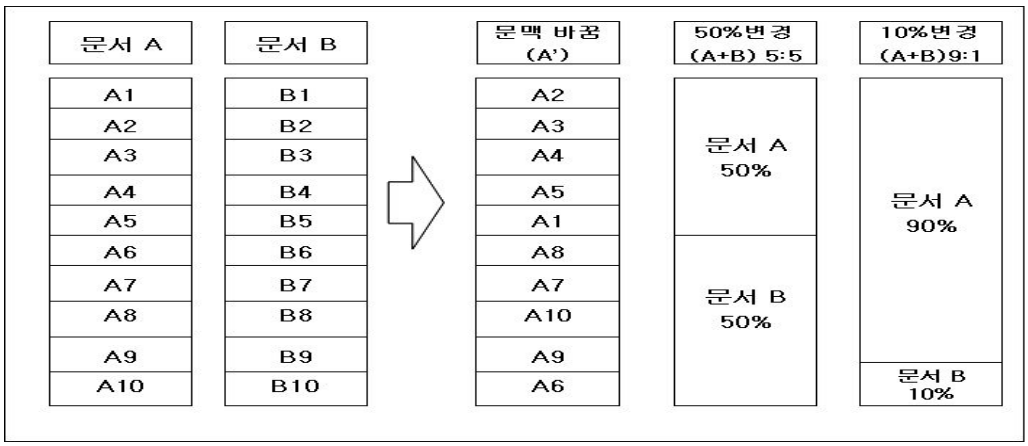


그림 4.5 두 번째 실험문서 집단의 문서 구성 형태

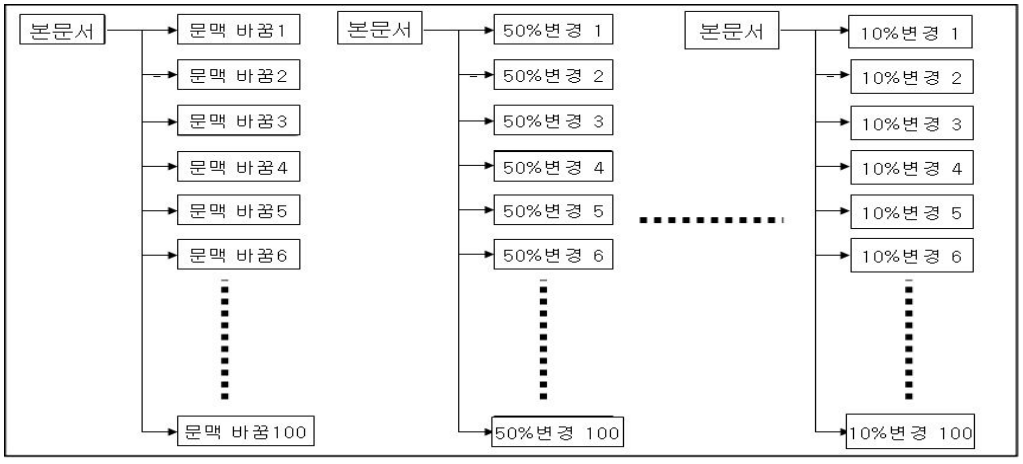
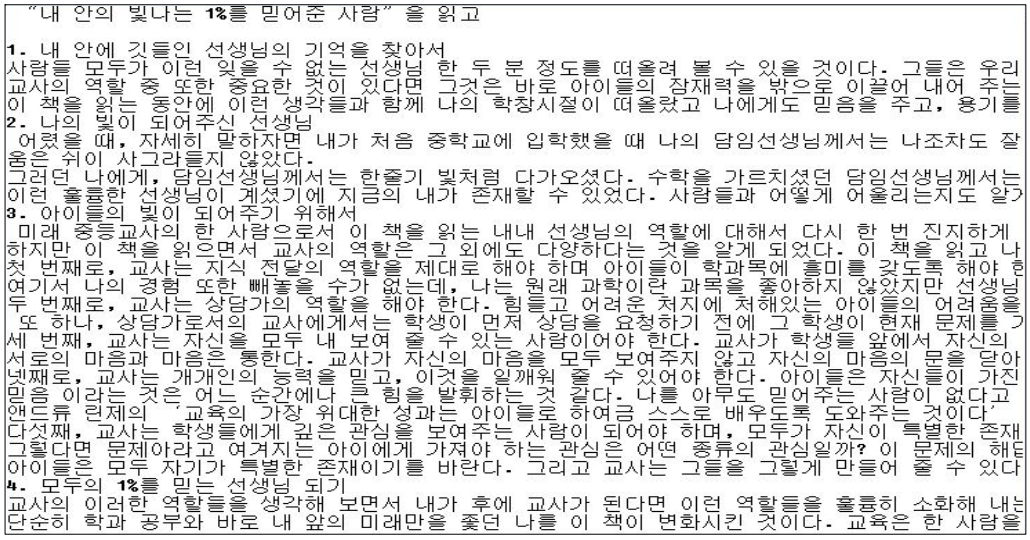
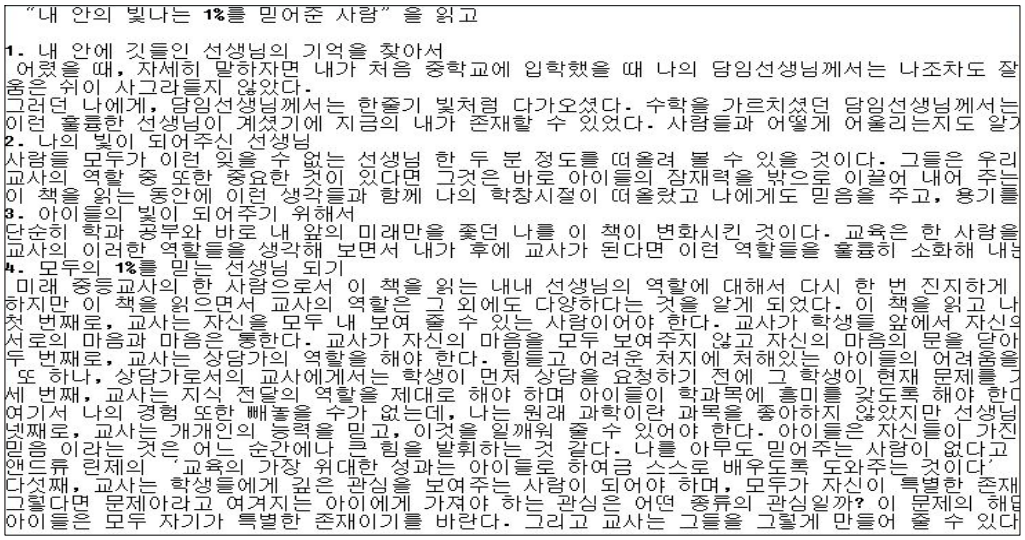


그림 4.6 두 번째 실험문서 집단의 문서 비교 방법


“내 안의 빛나는 1%를 믿어준 사람” 을 읽고

1. 내 안에 깃들인 선생님의 기억을 찾아서
사람들 모두가 이런 잊을 수 없는 선생님 한 두 분 정도를 떠올려 볼 수 있을 것이다. 그들은 우리
교사의 역할 중 또한 중요한 것이 있다면 그것은 바로 아이들의 잠재력을 밖으로 이끌어 내어 주는
이 책을 읽는 동안에 이런 생각들과 함께 나의 학창시절이 떠올랐고 나에게도 믿음을 주고, 용기를
2. 나의 빛이 되어주신 선생님
어렸을 때, 자세히 말하자면 내가 처음 중학교에 입학했을 때 나의 담임선생님께서는 나조차도 잘
움은 쉬이 사그라들지 않았다.
그러던 나에게, 담임선생님께서는 한줄기 빛처럼 다가오셨다. 수학을 가르치셨던 담임선생님께서는
이런 훌륭한 선생님이 계셨기에 지금의 내가 존재할 수 있었다. 사람들과 어떻게 어울리는지도 알
3. 아이들의 빛이 되어주기 위해서
미래 중등교사의 한 사람으로서 이 책을 읽는 내내 선생님의 역할에 대해서 다시 한 번 진지하게
하지만 이 책을 읽으면서 교사의 역할은 그 외에도 다양하다는 것을 알게 되었다. 이 책을 읽고 나
첫 번째로, 교사는 지식 전달의 역할을 제대로 해야 하며 아이들이 학과목에 흥미를 갖도록 해야 한
여기서 나의 경험 또한 빼놓을 수가 없는데, 나는 원래 과학이란 과목을 좋아하지 않았지만 선생님
두 번째로, 교사는 상담가의 역할을 해야 한다. 힘들고 어려운 처지에 처해있는 아이들의 어려움을
또 하나, 상담가로서의 교사에게서는 학생이 먼저 상담을 요청하기 전에 그 학생이 현재 문제를 가
세 번째, 교사는 자신을 모두 내 보여 줄 수 있는 사람이어야 한다. 교사가 학생들 앞에서 자신의
서로의 마음과 마음은 통한다. 교사가 자신의 마음을 모두 보여주지 않고 자신의 마음의 문을 닫아
넷째로, 교사는 개개인의 능력을 믿고, 이것을 일깨워 줄 수 있어야 한다. 아이들은 자신들이 가진
믿음 이라는 것은 어느 순간에나 큰 힘을 발휘하는 것 같다. 나를 아무도 믿어주는 사람이 없다고
앤드류 린제의 ‘교육의 가장 위대한 성과는 아이들로 하여금 스스로 배우도록 도와주는 것이다’
다섯째, 교사는 학생들에게 깊은 관심을 보여주는 사람이 되어야 하며, 모두가 자신이 특별한 존재
그렇다면 문제아라고 여겨지는 아이에게 가져야 하는 관심은 어떤 종류의 관심일까? 이 문제의 해답
아이들은 모두 자기가 특별한 존재이기를 바란다. 그리고 교사는 그들을 그렇게 만들어 줄 수 있다
4. 모두의 1%를 믿는 선생님 되기
교사의 이러한 역할들을 생각해 보면서 내가 후에 교사가 된다면 이런 역할들을 훌륭히 소화해 내는
단순히 학과 공부와 바로 내 앞의 미래만을 좇던 나를 이 책이 변화시킨 것이다. 교육은 한 사람을

그림 4.7 두 번째 실험문서 집단의 본문서의 실제 예


“내 안의 빛나는 1%를 믿어준 사람” 을 읽고

1. 내 안에 깃들인 선생님의 기억을 찾아서
어렸을 때, 자세히 말하자면 내가 처음 중학교에 입학했을 때 나의 담임선생님께서는 나조차도 잘
움은 쉬이 사그라들지 않았다.
그러던 나에게, 담임선생님께서는 한줄기 빛처럼 다가오셨다. 수학을 가르치셨던 담임선생님께서는
이런 훌륭한 선생님이 계셨기에 지금의 내가 존재할 수 있었다. 사람들과 어떻게 어울리는지도 알
2. 나의 빛이 되어주신 선생님
사람들 모두가 이런 잊을 수 없는 선생님 한 두 분 정도를 떠올려 볼 수 있을 것이다. 그들은 우리
교사의 역할 중 또한 중요한 것이 있다면 그것은 바로 아이들의 잠재력을 밖으로 이끌어 내어 주는
이 책을 읽는 동안에 이런 생각들과 함께 나의 학창시절이 떠올랐고 나에게도 믿음을 주고, 용기를
3. 아이들의 빛이 되어주기 위해서
단순히 학과 공부와 바로 내 앞의 미래만을 좇던 나를 이 책이 변화시킨 것이다. 교육은 한 사람을
교사의 이러한 역할들을 생각해 보면서 내가 후에 교사가 된다면 이런 역할들을 훌륭히 소화해 내는
4. 모두의 1%를 믿는 선생님 되기
미래 중등교사의 한 사람으로서 이 책을 읽는 내내 선생님의 역할에 대해서 다시 한 번 진지하게
하지만 이 책을 읽으면서 교사의 역할은 그 외에도 다양하다는 것을 알게 되었다. 이 책을 읽고 나
첫 번째로, 교사는 자신을 모두 내 보여 줄 수 있는 사람이어야 한다. 교사가 학생들 앞에서 자신의
서로의 마음과 마음은 통한다. 교사가 자신의 마음을 모두 보여주지 않고 자신의 마음의 문을 닫아
두 번째로, 교사는 상담가의 역할을 해야 한다. 힘들고 어려운 처지에 처해있는 아이들의 어려움을
또 하나, 상담가로서의 교사에게서는 학생이 먼저 상담을 요청하기 전에 그 학생이 현재 문제를 가
세 번째, 교사는 지식 전달의 역할을 제대로 해야 하며 아이들이 학과목에 흥미를 갖도록 해야 한다
여기서 나의 경험 또한 빼놓을 수가 없는데, 나는 원래 과학이란 과목을 좋아하지 않았지만 선생님
넷째로, 교사는 개개인의 능력을 믿고, 이것을 일깨워 줄 수 있어야 한다. 아이들은 자신들이 가진
믿음 이라는 것은 어느 순간에나 큰 힘을 발휘하는 것 같다. 나를 아무도 믿어주는 사람이 없다고
앤드류 린제의 ‘교육의 가장 위대한 성과는 아이들로 하여금 스스로 배우도록 도와주는 것이다’
다섯째, 교사는 학생들에게 깊은 관심을 보여주는 사람이 되어야 하며, 모두가 자신이 특별한 존재
그렇다면 문제아라고 여겨지는 아이에게 가져야 하는 관심은 어떤 종류의 관심일까? 이 문제의 해답
아이들은 모두 자기가 특별한 존재이기를 바란다. 그리고 교사는 그들을 그렇게 만들어 줄 수 있다

그림 4.8 두 번째 실험문서 집단의 본문서의 문단의 순서를 뒤바꾼 실제 예

3. 실험결과

1) 첫 번째 실험문서 그룹에 대한 실험 결과

첫 번째 실험문서 집단에 대한 유사도 측정 결과는 문장 단위로 추출된 명사들을 하나의 색인어로 보는 단어 그룹 색인 생성 방법이 가장 높은 유사도(100%)가 나오는 것을 알 수 있다. 이는 실험에 사용된 비교문서를 본 문서의 대한 크기만 조절하여 생성하였기에 결과적으로 분할된 비교문서의 전체가 본문서의 일부와 내용이 같기 때문이다. 그러나 단어 그룹 색인 생성 방법의 높은 문서간 유사도는 문장내의 일부 단어를 변경하거나, 수정하였을 때는 정확도가 현저하게 낮아지는 단점을 가지고 있다. 단어 그룹 색인 생성 방법을 제외한 다른 색인 생성 방법들을 살펴보면 단어 색인 생성 방법의 정확도가 68.41%로 단어쌍 색인 생성 방법의 평균 정확도 86.50%보다 문서간 유사도가 낮은 것을 확인 할 수 있었다. 이 실험에서는 문서간 유사도 측정 시스템뿐만 아니라 사람이 직접 문서간 유사한 정도를 살핀 결과 단어 색인 생성 방법에서는 같은 주제의 과제이기에 같은 단어의 사용이 많아서 실제로 유사하지 않은 문서에 대해서도 평균적으로 30%이상 비슷하다는 결과가 나왔지만 단어쌍 색인 생성 방법에서는 실제로 유사한 문서들간의 유사도만 높게 나오고, 유사하지 않은 문서에 대해서는 낮은 유사도가 나왔다. 이는 교수·평가자들이 같은 주제의 문서간 유사도를 측정할 때는 단어 보다는 단어쌍 색인 생성 방법으로 유사도를 측정하는 것이 평가의 신뢰도를 높일 수 있다는 결론이 나온다. 또한 단어쌍 색인 생성 방법에서 추출된 단어쌍의 개수가 유사도 측정 결과에 영향을 미치는 것을 알 수 있는데, 실험결과 슬라이딩 윈도우의 크기를 3으로 하였을 때 가장 높은 정확도

를 보인다. 한 문장 내에서 추출할 수 있는 모든 단어쌍을 추출한 가변적 슬라이딩 윈도우 색인 생성 방법의 실험 결과를 보면 단어쌍 색인 생성 방법 중에서 가장 낮은 정확도를 보인다. 이는 색인으로 추출된 단어쌍의 개수가 너무 많거나 또 너무 적으면 정확률이 낮아지며, 적절한 개수의 단어쌍을 추출하는 방법으로 슬라이딩 윈도우의 크기를 3으로 하였을 때 가장 높은 정확도를 보이는 단어쌍의 개수를 추출한다는 것을 알 수 있다. 그림 4.8은 각 색인 생성 방법에 따른 문서간 유사도 측정 결과이다.

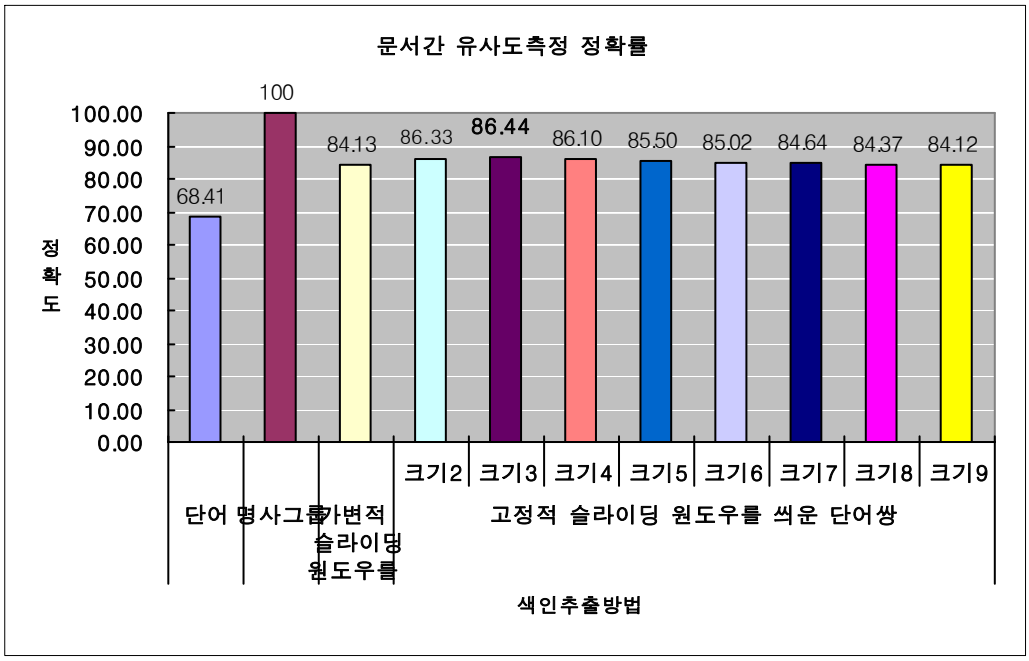
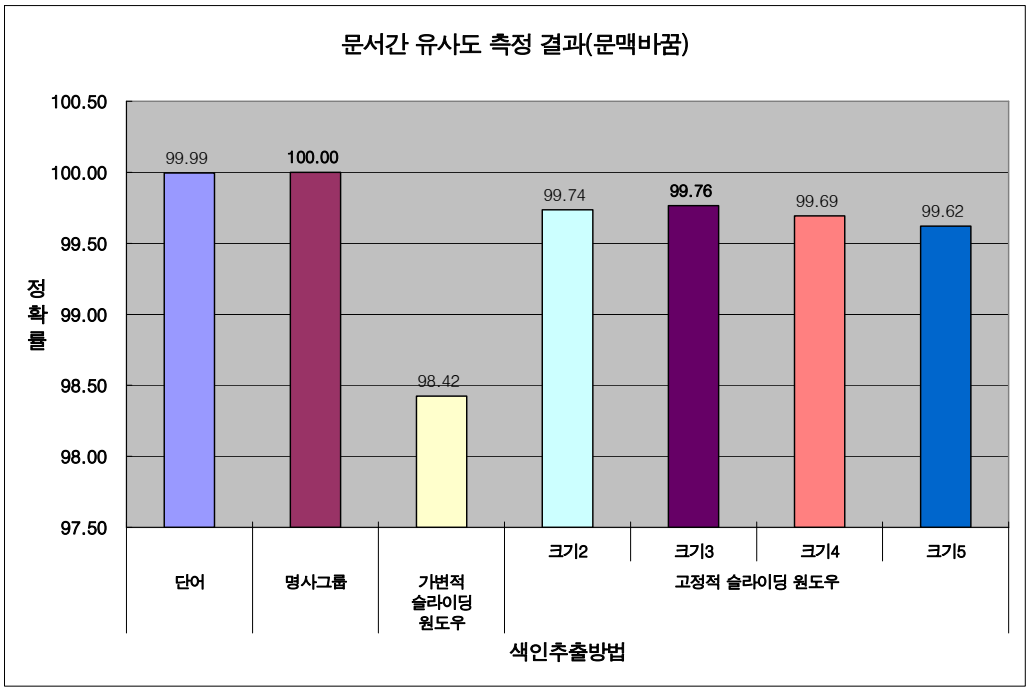


그림 4.9 첫 번째 실험문서 집단의 문서간 유사도 측정 결과

2) 두 번째 실험문서 집단에 대한 실험 결과

두 번째 실험문서 집단은 500개의 본 문서를 각각 3가지 유형으로 나누어 구성하였다. 이는 문서의 길이에는 변화를 주지 않고 문서의 내용만을 변경한 비교문서를 작성하여 본 문서와의 유사도 측정 시 결과가 어떻게 나오는지를 알아보기 위한 실험을 하기 위함이다.

첫 번째 유형은 본문서(A)안에서 단지 문단의 위치만을 바꾼 형태(A′)로 바꾼 문서와 이를 같은 유형의 다른 문서들과 비교하였더니 문서A와 문서A′의 비교에 대해서만 두 문서가 유사하다고 나오고 문서A와 문서B′,문서C′등의 비교에서는 두 문서가 유사하지 않은 것으로 나타났다.



[그림 4.10] 두 번째 실험문서 집단의 문서간 유사도 측정 결과(문맥 바꿈)

두 번째 유형은 본문서(A)의 90%는 그대로 두고 나머지 10%는 같은 주제의 다른 문서(B)에서 가져와 결합한 형태(AB)로 이를 같은 유형의 다른 문서들과 비교하였더니 결합한 문서(AB)와 본 문서(A)를 비교했을 때는 90%정도 유사한 것으로 나오고, 같은 주제의 다른 문서(B)와 결합한 문서(AB)를 비교했을 때 10%정도 유사하다는 결과가 나왔다.

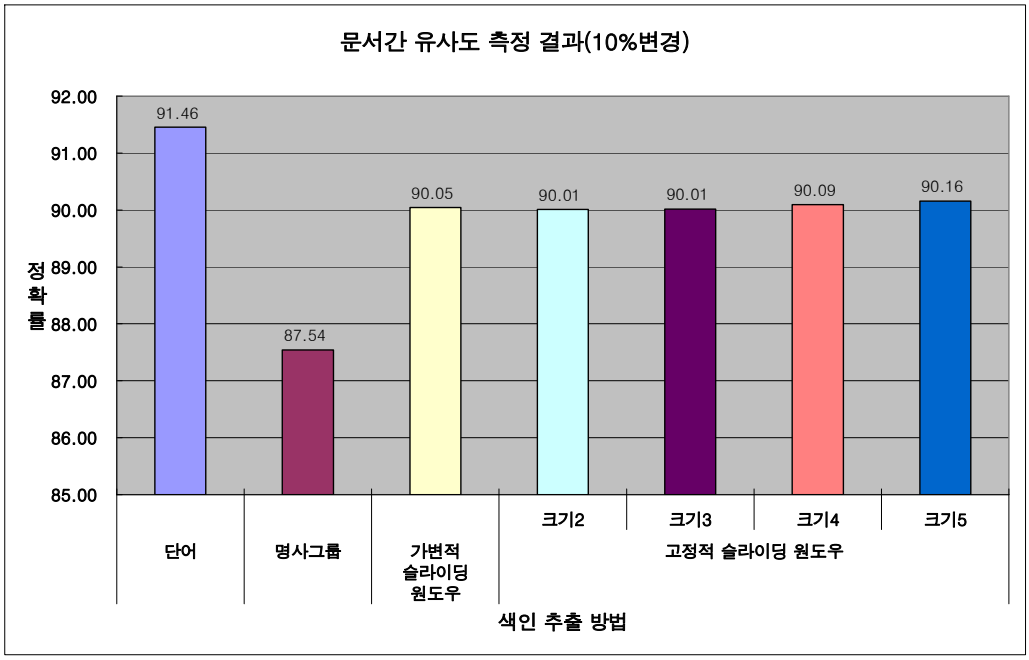


그림 4.11 두 번째 실험문서 집단의 문서간 유사도 측정 결과(10%변경)

세 번째 유형은 본문서(A)의 50%는 그대로 두고 나머지 50%는 같은 주제의 다른 문서(B)에서 가져와 결합한 형태(AB)로 이를 같은 유형의 다른 문서들과 비교하였더니 결합한 문서(AB)와 본 문서(A)를 비교했을 때와 같은 주제의 다른 문서(B)와 결합한 문서(AB)를 비교했을 때 각각 50%정도 유사하다는 결과가 나왔다.

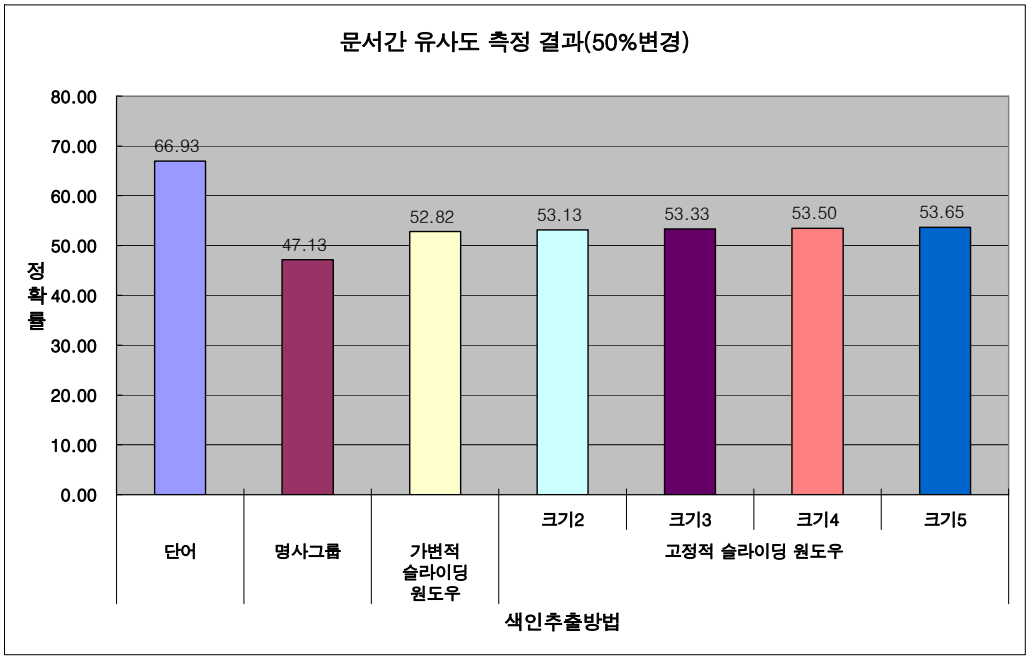


그림 4.12 두 번째 실험문서 집단의 문서간 유사도 측정 결과(50%변경)

Ⅴ. 결론

본 논문에서는 기존의 색인추출방법인 단어와 단어쌍 중에서 어떤 것이 정확도가 높으며, 또한 단어쌍 색인추출방법에서는 추출된 단어쌍 개수가 정확도에 어떤 영향을 미치는지 알아보기 위해 새로운 색인추출방법을 제시하여 유사도 측정 계산식에 따른 정확도를 실험·분석하였다.

그 결과 색인 추출 방법에서는 단어 색인 생성 방법의 정확도가 68.41%로 단어쌍 색인 생성 방법의 평균 정확도인 86.50%보다 18.09%더 낮은 정확률을 보여 단어 색인 생성 방법 보다는 문서의 문맥 정보를 포함한 단어쌍 색인생성 방법이 더 좋은 것을 알 수 있었다. 또한 단어쌍 색인 추출 방법 중에서는 슬라이딩 윈도우의 크기를 3으로 하였을 때 정확도가 86.44%로 단어 그룹 색인 생성 방법을 제외하고 가장 높은 정확도를 보이는 것을 알 수 있었다. 본 논문에서 제안한 단어 그룹 색인 생성 방법은 정확도가 100% 가장 높은 정확도를 보였으나, 문장 내에 변화를 주지 않고 본문서의 일부 혹은 전부를 그대로 복사하여 문서를 만들었을 경우에만 해당된 정확도이기에 본문서의 일부 혹은 전부의 문장내의 한 단어만이라도 변화를 준다면 정확도가 현저하게 낮아지는 단점을 가지고 있다.

본 연구를 통해 기대되는 효과로는 문서간 유사도 측정 시스템 구축시 사용할 여러 색인어 추출 방법들 중에서 유사도 측정 정확율 높일 수 있는 색인어 추출 방법을 제시하여 평가의 신뢰성을 높일 수 있다는 것이다.

그러나 좀 더 정확한 유사도 측정을 위해서는 명사만을 추출하여 유사도 측정을 할 것이 아니라 문서의 문맥을 이해할 수 있는 다른 방법을 찾아내어 문서간 유사도 측정의 정확도를 높일 필요가 있겠다. 또한 왜 슬라이딩 윈도우 크기가 3일때 가장 높은 정확률을 보이는지에 대한 보충 연구가 필요할 것이다.

참 고 문 헌

[1] 박수용, 서정연, 김학수, 고영중, “유사도 측정 기법을 이용한 효율적인 요구 분석 지원 시스템의 구현”, 정보과학회 논문지 제27권 제1호, pp13-23,2000.

[2] 김혜숙, “단어 및 단어쌍 별 빈도수를 이용한 문서간 유사도 측정”, 전남대학교 전산학과 석사학위논문. 2003.

[3] 이재윤, 최보영, 정영미, “문헌 자동분류에서 용어 가중치 기법에 대한 연구”,제7회 한국정보관리학회 학술대회 논문집, pp41-44, 2000.

[4] 조성용, “주제어 유사도 분석에 기반한 협업문서 생성제어 시스템 구현”, 전남대학교 대학원 석사학위논문, 2003.

[5] 정영미, “지식분류의 자동화를 위한 클러스터링 모형 연구, 한국정보관리학회지”, pp.203-230, 2001.

[6] 박수용, 서정연, 김학수, 고영중, "유사도 측정 기법을 이용한 효율적인 요구 분석 지원 시스템의 구현", 정보과학회 논문지 제27권 제4호, 2000.

[7] 심광섭, 인접 조건 검사에 의한 초고속 한국어 형태소 분석, 한국정보과학회 논문지 소프트웨어 및 응용, 31권 1호, pp.89-99, 2004. .

[8] 조일원, “유사 서술형 과제물 검사 시스템의 구현” 충북대학교 교육대학원 석사학위 논문, 2005

[9] Martin W.J.R., A1 B.P.F., and van Sterkenburg P.J.G., "On the processing of a text corpus: From textual data to lexicographic information," in Lexicography: Principles and Practice(Applied Language Studies Series), Hartmann R.R.K., Ed. London: Academic, 1983.

[10] Salton G. and McGill M.j., introduction to Modern Information Retrieval (Computer Series), New York:McGraw-Hill, 1983

[11] Ash R.. Information Theory, New York:Wiley-Interscience. 1965

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

[13] Jobbins A. and Evett L., "Text Segmentation Using Reiteration and Collocation," Proceedings of the COLING-ACL'98, 00.614-618, August 1998.

ABSTRACT

A Methodology for Semantic Similarity according to Index Extraction among Documents

Yun, Su Yeon
Major in Computer Science Education
Graduate School of Education
Sungshin Women's University

People can easily resolve problems through the mechanism 'Internet' and also can make things easier using various computer technologies. However, at the same time, people make a big deal social issue so called plagiarism by the fact that they can get thousands of pieces of information from the internet after a few minutes searching. Beacuse of this, assessors, and professors, are struggling for articulate evaluation of their students' work. To improve reliability of the evaluations, people use numerous index search and a Methodology for Semantic Similarity system through similarity measurement.

The dissertation which it sees the what kind of thing accuracy is high from in the word which is a indexed extraction method of existing and word pair, also the word pair modification which is

extracted what kind of effect should have gone mad to an accuracy, the hazard which it examines presents a new indexed extraction method from word pair indexed extraction method and test it analyzed the accuracy which also the beginning of history follows in measurement calculation. This test it leads and the indexed extraction method where the accuracy is high and the reliability against an evaluation volition evaluation result and it searches Professor inside to sleep it raises it does. The hazard 4 branch indexed extraction method which will reach (word index, the personage group, the size sliding window the word pair which puts on master drawing right (the size 2-9), variable sliding window the word pair which puts on master drawing right) the beginning of history it did a measurement test with smallest weight calculation.

The testing method divided impression of a book 10 thing of 10 page quantity which from the actual teachers' college unit hands over with subject with 1 page and create the data of 100 things which in the object which thing accuracy probably is higher from in word and word pair indexed extraction method, it examined, the indexed word meaning modification of word pair what kind of effect should have gone mad to an accuracy, against it it tested. The accuracy where word pair indexed extraction method is higher the result word than it seems, the indexed word meaning modification which word pair is extracted gives an effect to accuracy and the possibility of knowing the fact that the accuracy changes it was.