

자동초록 작성시에 발생하는 유사의미 문장요소들의 통합에 관한 연구*

A Study on the Integration of Similar Sentences in Automatic Summarizing of Document

이 태 영(Tae-Young Lee)**

목 차

- | | |
|---------------------|-----------------|
| 1. 서 론 | 4. 1 유사기준의 영향요인 |
| 1. 1 연구의 필요성과 목적 | 4. 2 유사기준의 조사분석 |
| 1. 2 연구 내용과 범위 및 방법 | 4. 3 유사도 특정모형 |
| 2. 연구 동향 | 5. 유사문장의 통합 |
| 3. 문장간의 유사 추정 | 5. 1 문장 대표 선정 |
| 3. 1 문장요소의 전체적 일치 | 5. 2 문장 대표 생성 |
| 3. 2 문장요소의 부분적 일치 | 6. 결 론 |
| 4. 유사 기준과 측정 | |

초 록

유사문장의 식별 및 통합을 위하여 문장의 구성성분, 품사, 절유형, 위치 등이 미치는 영향을 조사하고 유사도측정 공식과 통합방안을 모색하였다. 문법적 요인보다는 문장간에 일치하는 단어의 수가 유사성에 영향을 미치며 표제어와 기능절도 관여되었다. 문장간의 유사도 측정 공식은 설튼의 유사도 측정식과 코싸 인계수를 혼합하여 사용하였다. 유사문장들의 통합에서 절들의 대체 방법을 사용하였는데 앞으로는 단어들의 대체 방법으로 전환하여야 할 것이다.

ABSTRACTS

The effects of the Case, Part of Speech, Word and Clause Location, Word Frequency etc. were studied in discriminating the similar sentences of the Korean text. Word Frequency was much related to the discrimination of similarity and Tilt word and Functional Clause were little, but the others were not. The cosine coefficient and Salton' similarity measurement are used to measure the similarity between sentences. The change of clauses between each sentence is also used to unify the similar sentences into a representative sentence.

* 이 논문은 1998년 전북대학교 연구기반조성비의 지원으로 이루어졌음.

** 전북대학교 인문과학대학 문헌정보학과 교수
접수일자 2000년 5월 13일

1. 서론

1.1 연구의 필요성과 목적

현 시대에는 수없이 쏟아지는 문헌들을 일일이 다 읽어보고 필요한 정보를 수수하고자 하는 것은 거의 불가능한 일로 간주되고 있다. 따라서 본문의 대용물 역할을 하는 초록의 필요성이 무엇보다 강조되고 있으며, 해마다 발생하는 문헌량이 가히 천문학적 숫자이기 때문에 초록자들을 도와 줄 수 있는 기계적 작성, 즉 컴퓨터에 의한 자동초록 작성이 절실히 되는 시점에 와 있다. 자동초록 작성이 현실로 다가오면 백과사전이나 편람들도 컴퓨터가 여러 문헌들에서 같은 주제와 맥락의 정보들을 발췌하여 종합 정리함으로써 자동으로 작성할 수 있게 된다.

초록이 자동으로 작성되는 과정을 설명하면 다음과 같다. (1) 초록 모형(형식적인 구조)의 정립과, (2) 문장의 구문과 의미분석이 선결되고, (3) 분석된 각 문장들 간에 동일한 의미가 식별되어 통합되며, 그것들에서 (4) 중요한 의미를 갖는 문장들의 인식이 뒤따르고, (5) 초록 모형에 따라 이들 문장들의 개념이 정렬되며, 여기서 (6) 초록문을 생성시킨다. 그리고 (7) 위 과정에 필요한 시소러스, 사전, 그리고 기억구조에 대한 설계가 갖추어져야 한다. 이들 중 (1)의 모형은 그간의 연구에서, ① 연구의 목적과 관점, ② 방법론, ③ 획득된 결과, ④ 유추된 결론 등과 같은 사항들이 초록이 갖추어야 할 요소들로 규정되었다. (2), (7)도 관련 학문 분야(자연언어처리 등)에서 많은 연구가 진전되어 있으며, (6)의 문장생성 부문도 역시

관련분야에서 많은 연구가 이루어지고 있다. 그런데 (3), (4), (5)의 항목들은 이와 비교하여 볼 때, 상대적으로 연구가 취약하므로 앞으로 보다 많은 탐구가 이루어져야 할 것으로 생각된다.

본 연구에서는 초록문을 만들 때에 필요하면서 상대적으로 탐구를 더 요구하는 (3), (4), (5) 중 첫째 요소인 (3) '유사한 의미의 식별 및 통합'에 관하여 탐구함으로써 자동초록작성에 기여를 하고자 한다.

1.2 연구 내용과 범위 및 방법

유사문장의 통합에는 문장들 간에서 유사의미를 식별하고 나아가 유사한 문장이라고 판정할 수 있게 하는 범례나 준거규칙이 필요하다. 이런 범례나 준거규칙은 문장을 구성하는 각 부분(절, 구, 단어) 및 문법적 변수들(문장 성분, 품사, 용언어미변화 등)이 실제 문장들에서 어느 정도로 유사성에 기여하는가 하는 점의 해독을 통해 이루어진다. 이 해독을 위하여 문헌정보학 분야의 논문(전문 10페이지 내외) 중 무작위적으로 선정된 표본문헌(2개)에서 전술한 경우들의 예를 편출하여 시뮬레이션(대학생 이상의 질문대상자 15명에게 유사성의 정도에 대해 반복 설문한다)을 통해 분석하고 나아가 유사도 측정의 최적 모형을 도출하였다. 그리고 이 최적 모형을 가지고 표본의 서론, 본론(각 장별로 다시 세분함), 결론 등의 각 부분을 섭렵하면서 해당되는 각각의 유사문장들을 식별·발췌한다. 발췌된 유사문장들은 문장요소의 삭제와 대체를 통하여 하나의 대표문장으로 탄생된다. '삭제방법'은 유

사문장 중에서 중심문장을 선정하여 대표문장으로 삼는데 선택된 문장이 너무 길면 단어나 절들을 삭제하여 조정한다. '대체방법'은 절들간의 유사성을 점검하여 이 정보로 절간의 대체를 통해 대표문장을 생성시킨다.

2. 연구 동향

자동초록과정에서 나타나는 유사문장의 판별과 깊게 관련되는 유사문헌 식별에 대한 연구는 자동분류분야에서 1960년대 이래로 많은 연구들이 이루어져 왔다. 그 대표적인 경우로서 Salton과 그의 동료들이 있다. 그는 1968년 이래로 문헌유사성에 대한 연구를 많이 하였는데, 최근에는 문헌내의 문헌피이스(pieces)들의 연결성을 측정하기 위하여 각각의 문헌피이스 벡터들의 밀접성을 측정하는 유사도공식을 사용하였다. 그리고 문헌피이스 벡터에 포함되는 용어들의 가중치를 계산하는 공식(용어빈도와 역문헌빈도를 이용함)을 제시하였다 (Salton, Allen, Singhal 1996). 또한 Salton 등은 문헌들에서 의미하이퍼텍스트연결의 자동생성(automatic generation of semantic hypertext links)을 위하여 연구하는 과정에서 백과사전의 '스모킹(Smoking)'이란 항목에 기술된 단락(또는 문장집합)들 간의 유사성을 점검하여 연관성이 0.20 이상인 단락들을 연결시켜 단락연관지도를 만들었다. 유사성 측정에는 위의 유사도공식이 사용되었다. 그는 기능적으로 동일한 단락들을 텍스트 세그먼트(segment)로 규정하였고 단락들이 갖는 링크의 숫자에 의거하여 몇 개의 가장 중심적인

단락을 추출해낼 수 있다고 하였다(Salton et al. 1997).

또 다른 유사도 측정 연구로서 Karov와 Edelman은 단어의 명료성(disambiguation)에 대한 연구를 하며 단어와 문장간, 문장과 단어간, 문장과 문장간의 유사도 측정을 공식화하려고 노력하였다. 이들은 유사측정에 있어 가중치, 유사성, 친화성의 개념을 도입하였다(Karov and Edelman 1998). Moens 등은 SALOMON 프로젝트에서 '범죄혐의'와 '법정의견'의 양 텍스트의 단락들을 색인어의 벡터로 나타내었다. 그리고 텍스트의 각 단락의 유사도를 측정하기 위해 코사인계수를 사용하였으며(Moens, Uyttendaele, and Dumortier 1999) Schutze도 자동으로 단어의 의미를 식별하려는 연구에서 단어벡터의 유사도를 측정하는데 코사인계수를 언급하였다(Schutze 1998).

초록문장 구조 및 생성에 관련되는 연구로서 Radev와 McKeown은 SUMMONS를 제시하였는데 그들은 뉴스기사로부터 특출한 정보들을 발췌하고 이를 템프릿으로 저장하였으며 요약연산자를 사용하여 개념의 결합 및 중복 개념의 삭제를 하였다 (Radev and McKeown 1998). 한편 Enderes-Niggemeyer 등은 논문을 초록하는데 있어서 문헌을 구조적으로 서론, 본론(부문1(서론, 본론, 결론), 부문2(서론, 본론, 결론),...),결론으로 보고 그것들을 정리하여 논문(초록)의 변함없는 의미 전개구조인 "논제 문장, 방법, 결과, 결론"에 맞추어 초록을 만들려고 노력하였다 (Enderes-Niggemeyer, Sigel, and Maier 1995).

국내에서는 한국어 초록문장을 작성할 때

대두되는 부절과 부절, 부절과 주절사이의 연결에 적용되는 복합문의 주절과 부절 간의 결합문제, 즉 주절과 부절이 합쳐 복합문을 이룰 때 부절 용언의 접속어미의 변화에 따른 의미론적 기능에 대한 연구가 윤평현, 이상태 등에 의해서 수행되었다. 윤평현은 접속어미를 “접속관계, 양보관계, 대립관계, 목적관계, 결과관계, 인과관계”로 나누었고(윤평현 1992), 이상태는 “조건(면, 어야, 으수록), -되, {-는지, -는데, 으니} 등에 대해서 심도 있는 논의를 하였고 명제연결의미로 “열거, 선택, 반복, 제시, 시간, 조건, 양보관계”를 정리하였다(이상태 1995). 또한 이삼형은 단락에서 서술되는 문장들 간의 관계를 1)수집, 2)부가, 3)공제, 4)인과, 5)이유, 6)비교/대조, 7)상세화, 8)문제/해결, 9)초담화(meta-discourse)와 같이 설정하였다(이삼형 1994, 63-90).

3. 문장간의 유사 추정

3. 1 문장요소의 전체적 일치

문장들 간에 문장요소들이 전체적으로 일치하면 그 문장들은 유사하다고 판단하게 된다. 그러한 경우는 다음의 두 가지가 있다.

- 1) 어순에 관계없이 두 문장을 이루는 단어가 전부 동일할 경우이다.(이음동의어도 동일한 단어로 인정한다) :
- 2) 어순에 관계없이 준동의 명사구와 준동의 용언구를 포함하여 두 문장의 단어가 전부 동일할 경우이다.(이음동의어는 동일한 단어로 간주한다. 준동의 명사구와

용언구란 명사와 명사구 및 용언들의 집합에서 서로 의미가 일치한다고 인정한 각각의 명사구와 용언구를 뜻한다.) 준동의 명사구와 준동의 용언구의 예를 들면 다음과 같다.

가) 준동의 명사구

- (1) 명사는 동일하나 조사가 다를 때(구성성분이 동일한 경우) :
 - ①동물은 도망간다 <-> 동물이 도망간다
 - ②가슴 왼쪽을 타구에 맞아 <-> 가슴 왼쪽에 타구를 맞아
- (2) 명사 구나 복합어를 이루는 최소자립형태소들이 서로 위치전도 될 때 :
 - ①이회택 감독 <-> 감독이회택
 - ②이회택 감독 <-> 감독 이회택
- (3) 명사 복합어 중 한 최소자립형태소가 다른 단어로 대체되거나 생략될 때 :
 - ①주식 투자가 바뀌었다 <-> 주식 투자양상 : 투자방법이 바뀌었다
- (4) '의' 로 연결되는 구에서 관형 위치가 바뀔 때 :
 - ①수학의 철학 <-> 철학의 수학
- (5) 명사 뒤에 '적, 성, 형' 이 붙을 때 :

①감수 <-> 감수성

②이성 <-> 이성적

나) 준동의 용언구

- (1) 용언의 종결형 어미나 보조어구가 변화할 때 :
 - ①조작한다 <-> 조작하였다 <-> 조작할 수 있다 <-> 조작하게 만들었다
 - ②아름답다 <-> 아름다운 듯 하다 <-> 아름답고 싶다
- (2) 용언의 연결형 어미가 변화할 때 :
 - ①조작하여 <-> 조작한 후

- ②조작하면 <-> 조작하니(까)
- ③맞아(서) <-> 맞어(서)
- (3) 반대 용언일 때:
 - ①주가를 높였다 <-> 주가를 낮추었다
- (4) 동사 어간에 “하다, 되다, 시키다” 등이 결합될 때:
 - ①운동하다 <-> 운동되다 <-> 운동시키다 (이태영 1998)

3. 2 문장요소의 부분적 일치

문장들 간에 문장요소 구분단위인 절이 몇 개 일치하거나 문장이 포함하고 있는 단어의 많은 수가 의미상으로 비슷할 때에는 완전 유사는 아니나 두 문장이 준 유사하다고 유추할 수 있다. 그 경우를 적어 보면 다음과 같다.

- 1) 유사한 절간에 전체적 일치가 일어날 때이다.(단일문 포함)
 - 가) 주절과 주절의 전체적 일치 :
 - 3.1의 1), 2)와 같다.
 - 나) 주절과 부절의 전체적 일치 :
 - 3.1의 1), 2)와 같다.
- 2) 유사한 절간에 부분적 일치가 일어날 때이다.(단일문 포함)
 - 가) 주절과 주절의 부분적 일치 :
 - (1) 사용된 단어는 같으나 문장의 구성성분비 중 일부 성분이 생략되었을 때(문장을 이루는 성분인 주어, 목적어(간접, 직접), 보어, 술어, 부사어 등에서 임의의 성분이 생략되었을 경우이다.)
 - ①삼성은 자동차산업을 대우에게 양도하였다 <-> 삼성은 자동차산업을 양도되었다
 - (2) 문장의 구성성분비는 일치하나 사용된 단어

가 일부 다를 때 :

- ①개가 차도로 온다 <-> 개가 차도로 간다
- ②철수는 떡을 현미에게 주었다 <-> 철수는 감자를 현미의 집에 가져 갔다
- (3) 문장의 구성성분과 사용된 단어가 각각 부분적으로 일치할 때 :
 - ①차길진 팬이에요 <-> 차길진이란 사람을 어떻게 압니까?
 - ②가을에는 편지를 써 보내자 <-> 편지갈피에 가을 낙엽 한장을 놓자
- 나) 주절과 부절의 부분적 일치 :
 - 3.2의 2)-가)에 준한다 (이태영 1998).

4. 유사 기준과 측정법

4. 1 유사 기준의 영향요인

3장에서 기술된 문장들 간의 일치에 작용하는 요소로 상정할 수 있는 항목들은 (1)문장에 쓰인 단어의 종류와 수(체언, 용언, 부사 등등), (2)문장의 구성성분 및 한 성분안에 쓰인 단어의 수, (3)문장의 절 수와 용언의 어미변화, (4)단어들이 문장에 쓰인 순서, (5)주절과 부절 등이 유사성에 관여한다고 할 수 있다. 이러한 항목이 문장의 유사성을 판단하는데 결정요소로 작용하는가를 알아보기 위하여 다음과 같은 명제를 세웠다. ‘중동의 명사구’에 대한 것은 과거의 연구(이태영 1998)가 있었기 때문에 그것으로 대체한다.

- ㉠문장의 구성성분의 차이는 문장간의 유사성에 영향을 미칠까?
- ㉡문장 구성성분의 구성소(단어)가 많아질

수록 유사성에 영향을 미치는가?

- ㉔ 용언과 체언이 발휘하는 유사 식별력은 동등한 것인가?
- ㉕ 명사, 동사, 형용사를 제외한 부사, 수사, 대명사 등이 유사 식별에 영향을 미칠까?
- ㉖ 절의 용언어미가 문장간의 유사성에 영향력을 발휘할 수 있을까? (-하, -되 등 포함)
- ㉗ 문장안의 주절과 부절은 문장의 유사성에 있어 서로 대등한 영향력을 미치는가?
- ㉘ 두 문장에서 몇 개의 단어(종류)가 일치해야 유사하다고 판단될까?
- ㉙ 기능적인 역할을 하는 절도 문장간 유사성에 영향을 미칠까?

위와 같은 명제와 함께 전통적으로 초록후보문장 발체에 사용되는 표제어, 단서어들의 유사성에 미치는 영향력을 고려할 수 있다.

4. 2 유사 기준의 조사분석

유사성에 관련되는 요인을 조사하기 위하여 앞 절에서 설정한 영향요인들에 관련된 문장 표본들의 예를 부록과 같이 작성하여 15명의 대상자에게 설문하였다. 유사 정도의 표시는 100%, 75%, 50+%(50%가 조금 넘는다고 생각할 때), 50%, 25%, 0%로 나타내게 하였다. 설문결과에서 각 문항의 문장들이 유사하다고 인정한 경우는 문항의 대표값이 50+ %를 넘을 때이다. 즉 50+ %가 유사 판단의 한 계치가 되는 것이다. 문항의 대표값은 15명이 판정한 값 중에서 가장 많이 출현한 %값으로 정하였다. 4.1절에서 기술한 영향요인들을 위의 기준을 가지고 그 순서대로 조사분석하면 다음과 같다.

(1) 문장 구성성분의 차이 :

문장 구성성분의 차이는 아래의 유사측정 1(부록의 질문1)의 결과에서 보여지듯이 ㉔: ㉕, ㉔:㉖, ㉕:㉖가 모두 100%이므로 단어의 성분이 주어, 목적어 등으로 변하여도 의미소통에는 지장을 주지 않는 것으로 나타났다. 결국 구성성분은 유사성에 영향을 미치지 못한다.

유사측정1 : 예문1)에서 ㉔:㉕ = (100%), ㉔:㉖ = (100%), ㉕:㉖ = (100%)의 유사정도

예문1)

㉔ 열람목록은 이용자에게 도서관의 장서 전체를 보여 준다

㉕ 이용자는 열람목록으로 도서관의 장서 전체를 볼 수 있다

㉖ 도서관의 장서 전체는 열람목록에서 이용자한테 (로) 보여진다

(2) 문장 구성성분의 구성소(단어)의 다소 :

문장 구성성분의 구성소에서 유사한 단어가 많을수록 유사성은 높아진다. 유사측정2(부록의 질문3)의 ㉔:㉕가 50+%, ㉔:㉖가 50+ %인 반면에 유사측정3(부록의 질문6)의 ㉔:㉕는 50%, ㉔:㉖도 50%이다. 그러므로 구성성분의 구성소의 다소는 유사성에 영향을 준다고 할 수 있다.

유사측정2 : 예문2)에서 ㉔:㉕?(50+%), ㉔: ㉖?(50+%), ㉕:㉖?(75%)의 유사정도

예문2)

㉔ 참고봉사, 자동목록, 자동색인은 서지학을 필요로 한다

㉔참고봉사, 자동목록, 자동색인은 전산언어학을 필요로 한다

㉕참고봉사, 자동목록, 자동색인은 언어통계학을 필요로 한다

유사측정3 : 예 문 3)에서 ㉔:㉕?(50%), ㉔:㉕?(50%), ㉕:㉔?(75%)의 유사정도

예문3)

㉔자동색인은 서지학을 필요로 한다

㉕자동색인은 전산언어학을 필요로 한다

㉖자동색인은 언어통계학을 필요로 한다

(3) 체언과 용언의 차이 :

체언과 용언이 유사성에 서로 차별적으로 적용될 수 있는가하는 의문이 있다. 아래의 유사측정4(부록의 질문8)에서 ㉔와 ㉕의 유사정도가 0%이고 ㉖와 ㉗가 0%인 점으로 보아 “정치학, 경제학, 인용한다, 무시한다”가 서로 대등한 비중으로 작용하였다고 판단된다. 그리고 유사측정5(부록의 질문10)에서도 ㉔와 ㉕가 75%, ㉖와 ㉗가 역시 75%로 나타나 “문화, 정치, 가져왔다, 촉진한다”는 4단어가 서로 같은 무게로 작용되었다고 할 수 있다. 따라서 명사와 동사나 형용사간에 유사력의 차이가 없다고 사려된다. (부록의 질문9, 11, 12 참조)

유사측정4 : 예 문 4)에서 ㉔:㉕ = (0%), ㉖:㉗ = (50%), ㉕:㉔ = (0%)의 유사정도

예문4)

㉔정치학을 인용한다

㉕경제학을 인용한다

㉖경제학을 다룬다

㉗경제학을 무시한다

유사측정5 : 예 문 5)에서 ㉔:㉕ = (75%), ㉖:㉗ = (75%)의 유사정도

예문5)

㉔사회와 문화의 발전을 가져왔다

㉕사회와 문화의 발전을 촉진한다

㉖사회와 문화의 발전을 홍보한다

㉗사회와 문화의 발전을 촉진한다

㉘사회와 정치의 발전을 촉진한다

(4) 부사, 수사, 대명사의 영향 :

명사, 동사, 형용사를 제외하고 부사, 수사, 대명사가 문장의 유사성에 얼마나 영향을 주는가를 측정할 필요가 있다. 아래의 유사측정 6(부록의 질문23과 24)은 이 품사들의 유사력을 점검한 것이다. ㉔:㉕, ㉔:㉖, ㉕:㉖, ㉔:㉗, ㉕:㉗, ㉘:㉙의 유사정도를 살펴보면 모두 50%이상으로 부사, 수사, 대명사가 문장간의 유사성에 영향을 미치지 보다는 미치지 않는 쪽으로 생각해야 된다.

유사측정6 : 예문6)에서 ㉔:㉕ = (50+%), ㉔:㉖ = (50%), ㉕:㉖ = (50+%), ㉔:㉗ = (50%), ㉕:㉗ = (75%), ㉘:㉙ = (50%)의 유사정도

예문6)

㉔타 분야 문헌의 주제별 분포를 대체적으로 규명하였다

㉕타 분야 문헌들의 주제별 분포를 중점적으로 규명하였다

㉖그 문헌들의 주제별 분포를 중점적으로 규명하였다

㉗그것들의 주제별 분포를 중점적으로 규명하였다

㉘그것들의 주제별 분포를 첫째로 규명하였다

㉙그것들의 주제별 분포를 셋째로 규명하였다

(5) 절의 용언어미 변화 :

문장요소 단위의 하나인 절에서 용언이 변화를 일으킬 때 그것에 따라 의미도 변하는지를 분석할 필요가 있다. 아래의 유사측정7(부록의 질문13과 14)의 유사정도 내역을 분석하면 용언어미가 '-면'과 '-는데'처럼 어미의 속성이 다르지 않는한 유사성에 차이를 보이지 않는다. (부록의 질문15도 참조)

유사측정7 : 예문7)에서 ①:②=(100%), ①:③=(75%), ①:④=(100%), ②:④=(50%), ③:④=(75%), ⑤:⑥=(75%), ⑤:⑦=(75%), ⑥:⑦=(25%)의 유사정도

예문7)

- ① 온라인 목록은 이용자가 많은 접근점을 갖게 되어 도서내용을 잘 탐색할 수 있다
- ② 온라인 목록은 도서내용을 잘 탐색할 수 있게 많은 접근점을 갖는다
- ③ 온라인 목록은 이용자가 많은 접근점을 갖게 되면 도서내용을 잘 탐색할 수 있다
- ④ 온라인 목록은 이용자가 많은 접근점을 갖게 되며 따라서 도서내용을 잘 탐색할 수 있다
- ⑤ 역시 사회학은 사회과학적 성향이 강하며 경제학 및 정치학 분야 문헌을 선호하고 있음을 알 수 있다
- ⑥ 역시 사회학은 사회과학적 성향이 강하므로 경제학 및 정치학 분야 문헌을 선호하고 있음을 알 수 있다
- ⑦ 역시 사회학은 사회과학적 성향이 강해서 경제학 및 정치학 분야 문헌을 선호하고 있음을 알 수 있다
- ⑧ 역시 사회학은 사회과학적 성향이 강하면 경제학 및 정치학 분야 문헌을 선호하고 있음을 알

수 있다

(6) 문장안의 주절과 부절의 차이 :

아래의 유사측정8(부록의 질문16)의 결과에 의하면 주절과 부절은 문장간의 유사성에 대한 영향력이 없는 것으로 판단된다.

유사측정8 : 예문8)에서 ①:②=(50%), ①:③=(75%), ②:③=(100%)의 유사정도

예문8)

- ① 책자형 목록은 카드목록과 함께 사용되기는 하지만 도서관의 자료검색을 위해서는 속보성이 결여되어 있다.
- ② 속보성이 결여되어 있는 책자형 목록은 도서관의 자료검색에 문제점을 제기한다.
- ③ 책자형 목록은 속보성이 결여되어 있어 도서관의 자료검색에 문제점을 제기한다.

(7) 문장간의 단어 일치 수 :

두 문장에서 일치하는 단어의 숫자가 문장간 유사성에 영향을 미칠 수 있다. 아래의 유사측정9(부록의 질문17, 18)에서 보면 ⑧:⑨처럼 구성성분이 둘이며 또 그 둘이 단어(명사와 함께 쓰인 조사는 명사에 포함시켜 한 단어로 한다)가 하나씩이면 문장요소의 1/2이 일치해도 비유사로 간주된다. 그리고 ①:②, ①:④, ②:③처럼 구성성분(성분 당 단어가 한 개씩)의 2/3가 일치하면 유사쪽으로 간주된다. 유사측정10(부록의 질문19, 20)에 있는

유사측정9 : 예문9)에서 ①:②=(50%), ①:③=(0%), ②:③=(25%), ①:④=(50+%), ①:⑤=(50%), ②:⑥

= (50+%), ㉔:㉕ = (25%)의 유사정도

예문9)

- ㉑ 사서가 주제를 선정하였다
- ㉒ 사서가 책을 선정하였다
- ㉓ 이용자가 책을 선정하였다
- ㉔ 사서가 주제를 연구하였다
- ㉕ 주제를 선정하였다
- ㉖ 사서가 선정하였다
- ㉗ 주제를 규명하였다
- ㉘ 정보전달을 규명하였다

예에서는 구성 성분이 4개인 문장의 경우 ㉑: ㉒, ㉑:㉓, ㉑:㉔에서 보는 바와 같이 구성성분 (단어를 하나씩 포함) 중 3개가 일치하면 유사로 판정하고 ㉑:㉔, ㉒:㉓, ㉒:㉔와 같이 두 개가 일치하면 비유사라고 한다. 이러한 추세로 보아 6개 구성성분 중 5개가 일치하면 물론 유사하다고 유추할 수 있고 실제로 유사측정10의 ㉑, ㉒, ㉔를 비교한 것의 값을 보면 실제에서도 일치한다.

유사측정10 : 예문10)에서 ㉑:㉒?(50+%), ㉑: ㉓?(50+%), ㉑:㉔?(25%), ㉒: ㉓?(25%), ㉒:㉔?(25%), ㉑: ㉔?(50+%), ㉑:㉒:㉓?(75%), ㉑: ㉔:㉕?(75%), ㉒:㉔:㉕?(50+%)의 유사 정도

예문10)

- ㉑ 열람목록은 이용자에게 도서관 소장물을 보여준다
- ㉒ 색인은 이용자에게 도서관 소장물을 보여준다
- ㉓ 열람목록은 사서에게 도서관 소장물을 보여준다
- ㉔ 열람목록으로 사서가 도서관 소장물을 정리한다 (처리한다)

㉑ 카드목록은 다른 보완도구와는 별도로 그 내용이나 체제기능면에서 특별한 도구로서 발전되어 왔다

㉒ 카드색인은 다른 보완도구와는 별도로 그 내용이나 체제기능면에서 특별한 도구로서 발전되어 왔다

㉓ 책자목록은 다른 보완도구와는 별도로 그 내용이나 체제기능면에서 특별한 도구로서 발전되어 왔다

그러므로 문장간의 일치되는 단어 수는 유사성에 영향을 주며 비교가 되는 두 문장의 단어 수가 서로 증가되어 문장 간에 일치하는 단어 수가 증가 할수록 유사성이 높아지는 현상을 보이고 있다.(부록의 질문21, 22, 26도 참조) 두 문장이 유사하다고 판정될 때 판단의 근거로서 상호 일치하는 단어 수를 알아 볼 필요가 있다. 다시 말하면 유사 판단 기준을 일치하는 단어 수로 삼는 것이다.

(8) 기능적인 절의 역할 :

아래의 유사측정11(부록의 질문25)에서 기능적인 절은 “이표에 의해 유추할 수 있다”와 “그 표는 말해 주고 있다”이다. 이것들이 포함되어 비교되어진 유사값과 (7)의 유사측정10에서 ㉑, ㉒, ㉔의 비교값에 비추어 보면 기능절도 의미절 보다는 조금 약하지만 유사척도에 있어 역할을 하고 있다고 볼 수 있다.

유사측정11 : 예문11)에서 ㉑:㉒ = (50%), ㉑:㉓ = (50+%), ㉒:㉓ = (25%), ㉑:㉔ = (75%), ㉑:㉕ = (50%)의 유사정

도

예문11)

- ㉑정치학은 역시 사회과학적 성향이 높다는 것을 이 표에 의해 유추할 수 있다
- ㉒경제학은 역시 사회과학적 성향이 높다는 것을 이 표에 의해 유추할 수 있다
- ㉓정치학은 역시 법학과 관련이 높다는 것을 이 표에 의해 유추할 수 있다
- ㉔정치학은 역시 사회과학적 성향이 높다는 것을 그 표는 말해 주고 있다

(9) 표제어의 역할 :

제목에 출현한 의미어의 유사성에 대한 영향력을 파악하기 위하여 유사측정12(부록의 질문32)의 유사하다고 판정난 문장쌍을 본다. ㉑와 ㉒는 ㉑와 ㉓, ㉒와 ㉓보다 유사성이 높다고 할 수 없는데도 선택된 것은 제목의 표제어의 영향이라고 판단할 수 있다.

유사측정12 : 예문12)에서 유사 문자쌍은 a-b임

예문12) 제목: 사회학과 경제학의 의존도 비교

- ㉑표 1에 제시된 바와 같이 우리나라 경제학 분야의 타분야 문헌의존도는 6.30퍼센트이다.
- ㉒표 1에 제시된 바와 같이 우리나라 사회학 분야의 자분야 의존도는 80.34퍼센트이고 타분야 의존도는 19.66퍼센트이다.
- ㉓우리나라 정치학 문헌은 표 1에 제시된 바와 같이 자분야 의존도 85.64퍼센트와 타분야 의존도 14.36퍼센트를 기록하고 있다.

(10) 단서어의 역할 :

아래의 유사측정13(부록의 질문33)에서 ㉑와 ㉓가 유사하고 ㉑와 ㉒가 유사하다고

판단하면서 ㉑와 ㉒가 유사치 않다고 하는 것은 “그렇게 구명된 연구결과는”이란 단서어가 유사력을 발생시키지 못한 결과로 귀착될 수 있다.

유사측정13 : 예문13에서 유사 문자쌍은 a-b, a-c, a-f, d-e, c-d임

예문13)

- ㉑그렇게 구명된 연구결과는 우리나라 학술 정보시스템의 효율적 운영에 기여할 수 있을 것이다.
- ㉒전술한 몇가지 결론은 우리나라 학술정보시스템의 효율적 운영을 돕는 기초자료가 될 것이다.
- ㉓그렇게 구명된 연구결과는 일본의 학술 정보시스템의 효율적 운영에 기여할 수 있을 것이다.
- ㉔그렇게 구명된 연구결과는 우리나라 정치시스템의 효율적 운영에 기여할 수 있을 것이다.
- ㉕전술한 몇가지 결론은 우리나라 정치시스템의 효율적 운영을 돕는 기초자료가 될 것이다.
- ㉖결론적으로 보아 우리나라 학술 정보시스템의 발전적 운영에 혁신을 이루었다고 할 수 있을 것이다.
- ㉗결론적으로 보아 미국 학술 정보시스템의 발전적 운영에 혁신을 이루었다고 할 수 있을 것이다.

위의 결과들을 정리하면 다음과 같다. 문장의 구성성분, 용언과 체언, “부사, 수사, 대명사”, 절의 용언 어미변화, 주절과 부절은 유사성을 제고하는데 영향력을 별로 발휘하지 못하는 것으로 나타났다. 그리고 문장 구성성분의 요소(단어)의 수와 문장간 일치하는 단어의 수는 유사성에 영향을 주며 표제어와 기능절도 유사성에 영향을 미친 것으로 분석되어질 수 있었다. 따라서 유사성은 두 문장간의

일치하는 단어의 수에 따라 유사 또는 비유사로 구분할 수 있다.

4. 3 유사도 측정모형

유사도를 측정하는데 있어 Salton과 그의 동료들이 제시한 측정모형은 2단계로 구성되어 있다. 먼저 문헌피이스 또는 단락에 출현한 단어들의 가중치를 측정하는 식을 용어빈도와 역문헌빈도를 이용하여 만들었고 가중치로 표시되는 단어들로 이루어진 단락벡터들 간의 유사도를 측정하기 위하여 벡터유사함수(vector similarity function)를 제시하였다. 단어의 가중치 측정 식은 아래의 모형1과 같이 표현된다 (Salton, Allen, and Singhal 1996).

$$\text{모형1 : } W_{ik} = \frac{tf_{ik} \cdot \log(N/n_k)}{\sqrt{\sum_{j=1}^n (tf_{ij})^2 \cdot (\log(N/n_j))^2}}$$

(여기서 W_{ik} 는 문헌 i 에 출현한 용어 k 의 가중치이며 tf_{ik} 는 문헌 i 에 출현한 용어 k 의 빈도수이다. N 은 장서(전체 문헌을 뜻함)에 포함된 문헌의 수, n_k 는 용어 k 를 갖고 있는 장서 내의 문헌 수이며 t 는 벡터내의 용어의 수이다.)

그리고 설튼 등은 유사도 측정식을 고안하였는데 그것은 모형2와 같다.

$$\text{모형2 : } \text{sim}(Q_j, D_i) = \sum_{k=1}^L W_{jk} W_{ik}$$

(여기서 Q 는질문을, D 는 문헌을, W_{jk} 는 질문 j 에 출

현한 k 단어의 가중치, W_{ik} 는 문헌 i 에 출현한 k 단어의 가중치를 말한다.)

한편 Moens 등(Moens, Uyttendaele, and Dumortier 1999)과 H. Schutze(Schutze 1998)는 각각 텍스트의 단락과 단어벡터의 유사성을 측정하는데 모형3과 같은 코사인계수

$$\text{모형3 : } \text{corr}(V_i, W_i) = \frac{\sum_{i=1}^n V_i \cdot W_i}{\sqrt{\sum_{i=1}^n V_i^2} \cdot \sqrt{\sum_{i=1}^n W_i^2}}$$

(여기서 V_i 는 단락 또는 단어 벡터 V , W_i 는 단락 또는 단어 벡터 W 를 뜻한다.)

Karov와 Edelman은 단어의 명료성(disambiguation)에 대한 연구를 하며 단어와 문장간, 문장과 단어간, 문장과 문장간의 유사도 측정을 공식화하려고 노력하였다. 이들은 유사측정에 있어 가중치, 유사성, 친화성의 개념을 도입하였다(Karov and Edelman 1998).

본고에서는 설튼의 유사도 측정식과 코사인계수를 겸용하는데 문장간의 유사도측정에는 코사인계수를, 절간의 유사도측정에는 Salton의 유사도 측정식을 사용한다. Salton 등이 백과사전의 '스모킹'이란 항목에 기술된 단락(또는 문장집합)들 간의 유사성을 점검하면서 유사성이 0.20 이상인 단락들을 연결시켜 단락연관지도를 만들었는데 (Salton et al. 1997) 본고에서 절간의 유사성을 조사하여 유사하다고 판정받은 절들을 그의 유사도 측정식으로 계산하여 보니 약 0.1을 기록하였다. 한편 코사인계수로 산정한 문장간의 한계치는

0.66이었다.

5. 유사문장 통합

유사한 문장들이 한 곳에 모이면 이 문장들을 통합하여 이들 문장들이 내포하고 있는 의미가 대표적으로 표현될 수 있는 하나의 문장을 만드는 과제가 등장한다. 이러한 통합 및 문장생성은 초록작성의 한 부분으로 작용되는데 즉 어떤 유사한 문장군들이 표의하는 개념이 초록의 한 구성부분(문장)으로 선정되었을 때 이들 유사문장들의 통합된 의미를 문장형태로 표현하는 과정이 따르기 마련인 것이다.

유사문장들을 정리하여 한 문장으로 만드는 것에는 두가지 방법이 있다고 할 수 있다. 첫째는 여러 문장 중에서 대표될 만한 문장을 뽑아 그 문장으로 하여금 여러문장을 대변토록 하는 '문장 대표 선정' 방법이다. 그리고 이 경우 문장길이가 고려되어야 한다. 둘째로는 여러 문장들을 분해하여 그 분해된 부분으로부터 대표문장을 생성하여 내는 '문장 대표 생성' 방법이 있다.

5. 1 문장 대표 선정

이것은 유사문장군 중에서 한 문장을 대표로 선정하는 방법이다. 선택방법은 유사한 여러문장들 중에서 다른 유사문장에도 중복하여 출현한 문장요소들을 많이 갖고 있는 문장을 대표로 선출한다(나머지 문장들을 그 대표문장에 통합된 것으로 간주한다). 여기서 문장요

소의 중복출현 횟수는 문장을 구성하는 단위인 절별로 계산하거나 또는 단어별로 계산하는 두가지 방법이 있을 수 있는데 본고에서는 절별 계산을 우선으로 하고 그것이 여의치 않을 경우 단어별 계산을 하여 대표를 선정한다. 아래의 예문13을 보면 ㉠가 3중복이 2, 2중복이 2로 대표문장으로 선정된다(때에 따라 3중복이 2, 2중복이 3인 경우와 3중복이 3, 2중복이 1인 것과 같은 경우들이 나타날 수 있는데 이때에는 후자를 대표로 선정한다) 이와 같이 중

예문13)

- ㉠도서관의 이용자들은 카드목록보다는 온라인 열람목록을 선호하고 있다.
- ㉡도서관의 이용자들은 온라인 열람목록을 선호하고 있다.
- ㉢온라인 열람목록이 카드목록보다는 도서관 이용자들에게 더 이용된다.

복출현 횟수가 많은 문장을 대표문장으로 선정하는데 선정된 대표문장이 길이가 너무 길어 줄여야 할 필요가 있을 때에는 출현횟수가 적은 단어나 절부터 줄인다. 삭제요령은 될 수 있는대로 단어는 같은 구성성분 상에 있는 것을 대상으로 하고 절도 '-고, -고', '-며, -며' 등과 같이 같은 용언어미의 성질이 반복해서 출현하는 곳에서 하나씩 줄여 나간다.

1) 단일문의 삭제

단일문을 삭제할 때에는 문장의 구성성분을 기준으로 삼는다. "주어, 목적어, 보어, 술어"의 4가지 성분을 기본으로 하여 같은 성분에 속해 나열된 단어들을 삭제의 범위에 넣는다. 삭

제시 문장의 길이는 초록문장의 적정선인 25-27단어 내외에 맞춘다.

예문14)

전북의 전주, 완주, 진안, 장수, 김제, 부안, 정읍, 고창, 순창, 익산, 군산, 옥구, 무주 지역의 공공도서관에서 도서관 이용자들은 기존의 분류와 주제명 및 서명 카드목록보다는 온라인 열람목록을 더 선호하고 있다 (36단어)

예문14에서 보면 삭제의 대상은 “전주, 완주, 진안, 장수, 김제, 부안, 정읍, 고창, 순창, 익산, 군산, 옥구, 무주지역”이 일차로 떠오른다. 그런데 단일문의 의미통합에서 삭제는 좀처럼 일어나기 힘든 사항이다. 왜냐하면 단일문은 25-27단어를 넘기기가 무척 어렵기 때문이다.

2) 복합문의 삭제

복합문으로 이루어진 유사문장들에서 문장 대표 선정 방법이 적용되는 경우는 그 유사문장들이 유사도는 비록 낮지만 서로 비슷하다고 판정을 받았을 때이다. 이러한 문장들의 특성은 각 문장들의 절들 중에서 서로 높은 유사도를 갖는 절들을 발견하기가 쉽지 않다는 것이다. 이럴 경우 대표문장을 선택하게 되는데 선택후 문장의 길이가 적정선을 훨씬 넘었을 때는 절 삭제를 시도한다. 예문15에서 ㉠와 ㉡문장은 유사도가 낮지만 서로 유사하다고 판정을 받은 문장들이다. 그런데 ㉠는 50개의 단어가 출현하고(한 단어의 구별은 띄어쓰기와 품사를 기준으로 하였음) ㉡는 40개가 출현하였다. 그리고 양 문장의 절들 중에서 유사성이 큰 것들을 발견하기가 쉽지 않다. 따라서

대표문장을 선정하고 대표문장의 일부의 절들을 삭제할 수 밖에 없는 경우이며 그 절차는 다음과 같다.

예문15)

㉠㉡온라인 열람목록은 (㉢지역사회 주민 및 도서관 이용자의 복리와 검색 효율성을 촉진시키고) (㉣(정보를 획득하는) 작업을 끊임없이 향상시킬 수 있고), (㉤(필요한 때는 시간과 장소에 구애받지 않고) (정보를 획득할 수 있어야 하며)), (㉥(다른 형태의 목록과 비교하여) (유지비가 경감되게 고안되어야 한다.))

(*여기서 2중 괄호로 묶여진 절들은 각각의 절이긴 하지만 한 맥락의 절로 취급된 것이다.)

㉢㉣온라인 목록은 (㉦저자, 서명, 주제어 등의 접근을 통하여) (㉧목록정보를 빠르고 정확하게 검색할 수 있도록 하고) (㉨주거지, 직장, 교육기관과 같은 다양한 장소에서 정보획득이 가능하고) (㉩다양한 자료에 접근과 유지가 쉬어야 할 것이다.)

유사문장이 2개 출현하였을 때에는 3개 이상 출현하였을 때처럼 중복출현의 수를 가지고 대표문장을 선정할 수 없기 때문에 긴 문장을 대표문장으로 선택한다. 대표문장으로 선정된 ㉠는 적정선의 길이가 될 수 있도록 절 삭제를 시도한다. ㉠의 ㉦은 전체 절에 걸리는 주어 역할을하여 필수적인 성분이다. 그 이하의 일차 괄호안에 드는 절들은 선택의 대상이 되는 것으로 출현의 고저로 선정이 결정되어진다. 여기서 유사성 분석결과 ㉠의 ㉢과 ㉡의 ㉢, ㉠의 ㉣과 ㉡의 ㉣이 유사(약 50%), ㉠의 ㉤과 ㉡의 ㉤이 약 유사(약 25%)로 판정 받았으므로 먼저 ㉠의 ㉢과 ㉡의 ㉢, ㉠의 ㉣

과 ①의 ㉔이 출현 빈도수 2로 대표문장 구성의 일원으로 선택된다. 이들 당선된 주어와 절들의 단어수를 헤아려 보면 27개이다. 따라서 나머지 절들 중 ㉔의 ㉔이 그 다음 첨가 대상 절이지만 길이상 위의 두 절과 주어로 문장을 만든다. 그러면 “온라인 열람목록은 정보를 획득하는 작업을 끊임없이 향상시킬 수 있고 필요한 때는 시간과 장소에 구애받지 않고 정보를 획득할 수 있어야 한다”로 생성된다.

5. 2 문장 대표 생성

유사문장 통합 중 유사한 여러 문장들을 분해하고 그 분해된 부분으로부터 대표문장을 생성하는 방법을 모색하는 것이 본절의 주제이다. 여기에는 상당히 유사한 절(4.2의 (7)항으로 보아 유사정도가 75% 이상인 것)들로 구성된 복합문 형태의 유사문장들이 그 대상이며 통합 과정은 문장들을 절별로 분석하여 재조합하는 방법으로 이루어진다.

5. 2. 1 절 통합 조건분석

절간에 통합을 전제로 한 문장의 분석 방식이 고안되어야 한다. 다시 말해서 다른 문장에 있던 절을 자신의 문장 속에 포함시키거나 또 자신의 절과 대체시킬 때 삽입이나 대체가 가능한 여러가지 조건들이 있다.

1) 다른 문장들에서 출현한 절들이 한 문장으로 무리 없이 합쳐지기 위해서는 원래 문장들에서 절들 간에 가졌던 절 개념의 ‘표의상태’가 현재 문장에서도 서로 일치해야 한다.

예문16에서 ①의 “비교분석법과 통계적 방

법을 사용하여”가 ④의 ‘\$’ 기호 앞에 오려면 후행하는 절의 선행 체언부에 ‘주제별 분포’와 용언부에 ‘진단-비교분석’이나 ‘구명’이 있어야 한다. 즉 ①의 ‘\$’ 기호 뒤에 오는 절의 선행 체언 및 용언과 같은 종류내지 성질을 가진 단어들이 있어야 한다.

예문16)

④경제학, 사회학 및 정치학 문헌에 인용된 타분야 문헌의 \$ 정보전달매체별 분포 및 국가별 분포를 진단하고 비교분석하여 그 결과를 종합하였다.

⑤이렇게 조사된 타분야 문헌을 “비교분석법과 통계적 방법을 사용하여” \$ 주제별 분포, 전달매체별 분포 및 국가별 분포를 각각 구명하였다.

다시 정리하여 말하면 위 예문들에서 “주제별분포, 정보전달분포, 국가별분포”로 이끌어지는 절 앞에 어떤 문장에서 발생한 절이 올 수 있으려면 그 발생한 문장에서 그 절뒤에 “주제별분포, 정보전달분포, 국가별분포”의 체언이 하나 이상 있어야 하고 또 “진단-비교분석, 구명”과 같이 용언을 가진 절이 결합되어 있어야 한다. 또한 체언부에서 “주제별분포, 정보전달분포, 국가별분포” 보다 하위개념어나 이음동의어가 대체되어 있어도 결합이 가능하다. 그것은 ‘사람’에 적용되는 내용은 “어린이, 청년, 노인 등”에도 다 적용되어질 수 있기 때문이다. 그러므로 절의 표의 하는 바를 나타내는 방식과 결합시 운영하는 방법을 마련하여야 한다.

2) 다른 문장들에서 출현한 절들이 한 문장으로 무리 없이 합쳐지기 위해서는 원래 문장

들에서 절들 간에 가졌던 용언어미의 성질이 같은 종류이어야 한다.

아래의 예문17에서 ㉔문장의 “사회학, 경제학, 역사학, 철학의 순위를 보이면서”가 ㉕문장의 “경제학, 사회학, 정치학 및 역사학에 집중되는 경향을 보이고 있으며”와 대체되어 ‘\$’ 표식 앞에 올 수 있다. 이것은 대체되는 절의 용언어미가 ‘며’와 ‘면서’ 이어서 용언어미 성질상 같은 부류에 속하며 또한 후행하는 절 체언이 각각 “타학문분야”와 “타분야”로 같은 뜻이기 때문에 가능하다.

예문17)

㉔우리나라 정치학 문헌의 주제별 분포는 “사회학, 경제학, 역사학, 철학의 순위를 보이면서” 적어도 14개 의 타학문분야에 의존하는 성향을 보이고 있다.

㉕우리나라 정치학자들이 인용하는 타분야 문헌의 주제는 “경제학, 사회학, 정치학 및 역사학에 집중되는 경향을 보이고 있으며” \$ 타분야 의존도

는 학문분야에 따라 서로 다르다.

본고에서는 절간의 관계를 윤평현(윤평현 1992)과 이상태(이상태 1995)의 연구를 참조하여 <표 1>과 같이 규정하였다

3) 다른 문장들에서 출현한 절들이 한 문장으로 무리 없이 합쳐지기 위해서는 원래 문장들에서 절들 간에 가졌던 절 개념들의 전개 상황을 알아야 한다.

아래의 예문18에서 ㉔의 ‘\$’ 기호 뒤에 오는 “비교분석법과 통계적 방법을 사용하여”가 ㉕의 ‘\$’ 기호 뒤에 오는 “인용문헌분석을 통하여”와 대체되려면 ㉔와 ㉕문장에서 ‘\$’ 기호 전후에 있는 절들의 전개적인 상황이 같거나 비슷해야 한다. ㉔와 ㉕문장에서 전문장을 살펴보면 각기 “타분야 문헌을 코드화하고”, “정보원을 조사하고”인데 각기 ‘문헌’을 나타내며 용언어미도 ‘고’로 되어 있어 문맥 상황이 비슷하다. 그리고 후문장을 살펴 보면

<표 1> 절 용언어미의 관계코드와 의미

코드: 관계 : 용언어미	
A	: 조건 : 거든, (더)ㄴ들, ㄹ진대, 면(단순)
B	: 조건 : 어야(필연적), 을수록(비례), ㄹ수록, 사록, 을지라도
D	: 대립 : 나, 아도, 지만, 나마, 건만, 려만
E	: 목적 : 려, 려고, 고자
F	: 결과 : 게, 게끔, 게시리, 도록
G	: 인과 : 아서, 니까, 므로, 느라고, 아
H	: 설명 : 되, 는바, 노니, 을새
I	: 양보 : 아도, 더라도, ㄹ지라도, ㄴ한들, ㄹ지언정
J	: 판단불명확의문제기 : 는지(배경-전경), 는데(배경-전경(춧점)),으니(전경-전경)
K	: 명제나열(동시발생) : 며, 고, 면서
L	: 상황전개 : 어, 여
M	: 선후관계 : 자, 다가

각기 “타분야 문헌의 분포를 구명하였다” “그들의 인용문헌의 분포를 구명하려”와 같이 되어 있어 표의상태가 “문헌분포-구명”으로 일치한다. 따라서 ㉔문장의 “비교분석법과 통계적 방법을 사용하여”와 ㉕문장의 “인용문헌분석을 통하여”는 상호의 위치에서 서로 대체될 수 있다.

예문18)

- ㉔사회과학자들의 문헌에 인용된 “타분야 문헌을 코드화하고” § “비교분석법과 통계적 방법을 사용하여” “타분야 문헌의 분포를 구명하였다”
- ㉕우리나라 사회과학자들의 “정보원을 조사하고” § “인용문헌분석을 통하여” “그들의 인용문헌의 분포를 구명하려”는 것이 이 연구의 목적이다

이러한 문장내의 절들의 전개정보를 표현하기 위해서는 앞 뒤 절간의 관계설정이 명확히 구명되어야 한다. Enderes-Niggemeyer 등은 텍스트표현을 위하여 관계를 사용하였다. 그들이 제시한 관계는 “제시(Presentational)-연속(Sequence), 개요(Schema)-부분(Part), 지원(Candidate), 공헌(Elaboration), 반복(Restatement), 예증/예(Exemplified-by/Example-for), 해결(Solutionhood), 부각(Highlighting), 원인/결과(Cause/Result)” 등과 같았다. (Enderes - Niggemeyer 1995) 또한 이삼형은 단락에서 서술되는 문장들 간의 관계를 1) 수집-공통된 의미의 나열, 2) 부가-중심되는 의미에 다른 의미를 덧붙임, 3) 공제-큰 의미에서 그 부분의 의미를 공제, 4) 인과-원인 사건의 선행과 결과 사건의 후행, 5) 이유-뒤의 문장의 의미가 앞 문장의 의미에 이유나 근거를 제공, 6) 비

교/대조-두 의미를 비교와 대조, 7) 상세화-일반적 진술과 구체적 진술이 결합(①셋트-멤버, ②일반-예시, ③전체-부분, ④과정-단계, ⑤대상-특질), 8) 문제/해결-문제를 제기하고 문제를 해결하는 방식, 9)초담화(meta-discourse)-내용의 전환 역할과 같이 설정하였다. (이삼형 1994, 63-90)

본고에서는 위에 나열된 관계 중 “도입(제시), 반복, 문제, 해결, 부각, 원인, 결과, 이유, 비교, 부각, 과정, 초담화”에 ‘마감’을 추가하여 절간의 관계설정으로 사용한다. 특히 ‘도입’은 “---을 위하여, --대한, 주어를 수식하는 절 등”과 같이 정하며 ‘마감’은 주절이 앞에 나열된 관계에 속하기 어려울 때 사용한다.

이상과 같이 유사 문장들을 통합시킬 때의 필요한 조건들을 (1) 절의 표의상태: 시간, 장소, 식물-성장, 곤충-생태 등, (2) 절의 용언어미 역할, (3) 절의 전개 상황으로 분석하였다.

5. 2. 2 절 조건 표현식

절들이 통합을 전제로 해서 문장내에서 갖고 있는 조건을 표현하고 또 보존하는 형식으로 Radev와 McKeown이 제시한 템프릿(Template)을 유용하게 사용할 수 있다. Radev와 McKeown은 SUMMONS를 제시하였는데 SUMMONS는 입력기사들의 특출한 사항들을 갖고 있는 템프릿의 집합으로부터 요약을 생성하고 있다. 한 뉴스기사로부터 특별한 정보의 조각들을 발췌하여 저장하는데 이 템프릿은 메시지해독시스템에 의해 생성된다. SUMMONS에서 채택하고 있는 MUC-4시스템의 템프릿과 SUMMONS의 것을 소개하면 <그림 1>과 같다. MUC-4는 테러리스트 영

MUC-4	SUMMONS
MESSAGE: ID TST3-MUC4-0010	MESSAGE; ID TST-REU-0001
MESSAGE: TEMPLATE 2	SECSOURCE:SOURCE Reuters
INCIDENT: DATE 01 NOV 89	SECSOURCE: DATE March 3,
LOCATION ELSALVADOR	1996 11:30
INCIDENT: TYPE ATTACK	PRIMSOURCE : SOURCE
.	INCIDENT: DATE March 3, 1996
PREP: INCIDENT TERRORIST	INCIDENT: LOCATION Jerusalem
CATEGORY ACT	INCIDENT: TYPE Bombing
PREP:INDIVIDUAL ID"TERRORIST"	HUM TGT: NUMBER "killed: 18"
.	PREP: ORGANIZATION ID
PHYS TGT: ID	
PHYS TGT: TYPE	
.	
HUM TGT: NAME	
HUM TGT: DESCRIPTION "1 CIVILIAN"	

<그림 1> MUC-4와 SUMMONS의 템프릿 표본

역을 운영하며 ‘저지름’, 희생자, 사건유형과 같은 필드를 갖고 있는데 템프릿 당 25개 필드가 있다(Radev and McKeown 1998).

본고는 문장과 절 템프릿을 동시에 운영한다. 절의 템프릿 제목은 가칭 “분포-분석”과 같은 절의 표의상태로 명칭하고 필드는 실제의 예들을 정리하여 담아 참고시킨다. 여기에는 이음동의어, 같은 맥락의 동일 성절어, 상하의어 등이 적힌다. 그 예는 다음의 <그림 2>와 같다. 한편 문장의 템프릿 제목은 마감절의 표의상태로 하고 여기에는 절 템프릿의 제목과 문장출현수, 포함되는 절, 위치 등이 적힌다. 그 예는 <그림 3>에 나와 있다.

5. 2. 3 통합 진행 절의 선정

5. 1 절에서 문장 대표 선정을 위해 유사의

미 문장들의 통합을 진행시키는 과정에서 대표문장의 선택은 대표 후보문장 내의 절들이 여러 문장들에서 중복하여 출현한 횟수의 다소를 가지고 결정한다고 하였다. 여기서도 대표문장에 포함되는 절이 되기 위해서는 그 절이 속하는 유사문장군 내의 각 문장들에서 얼마나 많이 중복 출현하였는가를 가지고 결정한다. 이와 같이 많이 출현한 순서대로 절들을 발췌하여 초록문장 길이의 적성선까지 골라 통합을 진행시키는데 경우에 따라 발췌 대상 마지막 순서에서 빈도수가 같은 절들이 복수로 있으면 초록문장의 짧막성에 비추어 짧은 길이의 절을 우선으로 한다.

그리고 선택된 유사의미 절들 중에서 어느 것을 선정하느냐는 문제는 각 절들 안에포함되어 있는 주제 단어들 중 많은 출현 빈도수

subject : 표의상태 components : 질출현수 components : 중복출현단어수 components : 출현단어수 instance : 질의 상위 예 instance : 질의 하위 예 instance : 동의 관계 예 essential mark field : 용언어미 성질 essential mark field : 전개상황-전 essential mark field : 전개상황-후 location : 절위치 temporary field :
--

〈그림 2〉 절 Template 예

subject : 표의상태 components : 문장출현수 components : 포함절 pattern : 주어+부절+술어 pattern : 부절+주어+부절+술어 pattern : 부절+부절+주절 location : temporary field :
--

〈그림 3〉 문장 템프릿의 예

를 갖고 있는 단어를 많이 포함하고 있는 절을 선정한다. 아래의 예문19를 보면 ㉠의 ㉡과 ㉢의 ㉣이 유사하고 ㉠의 ㉡과 ㉢의 ㉣ 및 ㉢의 ㉣이 유사하며 또 ㉠의 ㉡과 ㉢의 ㉣이 유사하다. 그리고 ㉠의 ㉤, ㉢의 ㉤, ㉢의 ㉤이 주어의 자격으로 합류된다. 여기서 유사 절군에서 하나의 대표 절을 선택하는 과정이 상기한 기준으로 전개된다. ㉠의 ㉤, ㉢의 ㉤, ㉢의 ㉤은 전부 동일한 단어이므로 어느 것이 선정되어도 상관이 없다. 다음의 ㉠의 ㉡과 ㉢의

㉣은 많이 출현한 주제어를 계산하면 ㉢의 ㉣이 '목록 5회', '검색 3회'인 반면에 ㉠의 ㉣은 '정보 5회', '획득 2회'이다. 따라서 ㉢의 ㉣을 대표절로 선정하며 ㉠의 ㉡과 ㉢의 ㉣ 및 ㉢의 ㉣에서는 ㉢의 ㉣이 선정된다. ㉠의 ㉡과 ㉢의 ㉣에서는 ㉠의 ㉡이 선택된다. 그러나 ㉠의 ㉡은 ㉢의 ㉣과 ㉢의 ㉣에 비해 중복출현 및출현 단어수에서 적으므로 대표문장의 일원이 되기는 어려워 보인다.

예문19 (예문15와 같은 문장임)

- ㉠ 온라인 열람목록은 (㉠지역사회 주민 및 도서관 이용자의 복리와 검색 효율성을 촉진시키고 (㉡정보를 획득하는) 작업을 끊임없이 항상시킬 수 있고), (㉢(필요한 때는 시간과 장소에 구애받지 않고) (정보를 획득할 수 있어야 하며)), (㉣(다른 형태의 목록과 비교하여) (다양한 자료에 쉽게 접할 수 있어야 한다.))
- ㉡ 온라인 목록은 (㉠저자, 서명, 주제어 등의 접근을 통하여) (㉡목록기입을 빠르고 정확하게 검색할 수 있도록 하고) (㉢주거지, 직장, 교육기관과 같은 다양한 장소에서 정보획득이 가능하고) (㉣지역적, 국가적, 국제적으로 다양한 자료에 접근이 가능하도록 하여야 할 것이다)
- ㉢ 온라인 목록은 (㉠필요한 때에 어디서나 정보를 검색할 수 있게) (㉡네트워크가 구축되어야 한다)

5. 2. 4 통합문장의 생성

통합문장에 포함될 절들이 선정되면 그 절들을 재조합하는 과정이 뒤따른다. SUMMONS에서는 현재 8개의 요약연산자와 200개의 입력 템프릿을 가지고 운영되는데(Radev and McKeown 1998, 478-479) 이 중에서 한 예를 들면 <그림 4>에서 보는 바와 같이 요약조건 및 과정을 적절히 표현하고 있다고 할 수 있다.

본고에서는 유사문장들의 대표문장을 생성하기 위해 5개의 연산자를 가동한다. 먼저 절간의 통합이 가능한가를 물어보는 통합가능연산자와 통합이 가능하였을 때 어떤 절로 통일시킬 것인가를 판정하는 절선택연산자, 그리고 통합문장을 생성하는 절조합 연산에 A와 B와 C형이 그것이다. 첫 번째의 통합가능연산

```
((#TEMPLATES == 2) &&
(T[1].INCIDENT.LOCATION == T[2].INCIDENT.LOCATION) &&
(T[1].INCIDENT.TIME < T[2].INCIDENT.TIME) && ...
(T[1].SECSOURCE.SOURCE != T[2].SECSOURCE.SOURCE)) ==>
(apply ('contradiction', 'with-new-account', T[1], T[2] ))
```

<그림 4> 부정(모순)연산자 규칙

```
CT(i)/ST(i) :: CT(j)/ST(j)
((CT(i) == CT(j) : 유사도) &&
(CT(i).표의상태 == CT(j).표의상태) &&
(CT(i).용언어미성질 == CT(j).용언어미성질) &&
(CT(i).전.전개상황 == CT(j).전.전개상황) &&
(CT(i).후.전개상황 == CT(j).후.전개상황)) ==>
(통합가능 CT(i).ST(i) :: CT(j).ST(j))
```

<그림 5> 통합가능 연산자

CT(i)/ST(i) :: CT(j)/ST(j)
 (((CT(i).출현수 > CT(j).출현수) ++
 (CT(i).중복출현단어수 > CT(j).중복출현단어수)) &&
 (CT(i).출현단어수 < CT(j).출현단어수) &&
 (적정단어수 > ST(u))) ==>
 (절선택 CT(i))

<그림 6> 절 선택연산자

CT(i)/ST(i) :: CT(j)/ST(j)
 ((CT(i).ST(i).절위치 < CT(j).ST(i).절위치) &&
 (CT(i).ST(i).절위치 > CT(j).ST(j).절위치) &&
 (주어.주절.절위치 < CT(i).절위치) &&
 (마감절.주절.절위치 > CT(j).절위치)) ==>
 (절위치 (주어.주절 + CT(i).ST(i) + CT(j).ST(j) + ..))

<그림 7> 절 조합연산자 A형

자에 들어가는 논리 항목들은 아래의 <그림 5>와 같이 “①절 간의 유사도, ②절의 표의상태, ③절의 용언어미 성질, ④절의 전개상황”이 있다.

두 번째의 절선택연산자(그림 6참조)에는 논리항목이 “①절의 출현수, ②절안에 다른절에도 중복 출현한 단어의 수, ③짧은 절(절안의 출현한 단어수), ④초록문의 적정한 단어수”가 포함된다.

세 번째, 절을 조합시켜 문장을 만드는 절 조합연산자는 ‘절의 위치’ 항목을 가지고 ST(u)의 절 순서를 정해 준다. 그런데 절조합 연산에는 3가지 경우를 고려해야 한다. 한국어 문장에서 “주어 + 부절 + ... + 마감절”, “부절 + 부절 + .. +주절”, 그리고 “부절(수식절) +

주어 + 부절 + ... + 마감절”과 같은 문장형태가 많이 있으므로 이 세가지 경우에 따른 절 조합 연산방식을 마련하였다. 첫번째인 A형을 소개하면 아래의 <그림 7>과 같다. 그리고 이 규칙에 의해 앞절의 예문19의 대표문장은 “온라인 열람목록은 목록기입을 빠르고 정확하게 검색할 수 있도록 하고 필요한 때에 어디서나 정보를 검색할 수 있다”와 같이 생성된다.

6. 결 론

자동으로 초록을 작성할 때에 필요하면서도 상대적으로 탐구를 더 요구하는 부분 중의 하나인 ‘유사한 의미의 식별 및 통합’에 관한

연구를하였다. 문장간의 유사성의 식별행태을 알아보기 위해 표본에서 유사하다고 인정되는 문장들을 묶어서 설문하였다. 그 결과 문장 간의 유사성을 판별할 때 문장의 구성성분, 용언과 체언, “부사, 수사, 대명사”, 절의 용언 어미 변화, 주절과 부절은 유사성을 제고하는데 영향력을 별로 발휘하지 못하는 것으로 나타났다. 그리고 문장 구성성분의 요소(단어)의 수와 문장간 일치하는 단어의 수는 유사성에 영향을 미치며 표제어와 기능절도 유사성에 영향을 미칠 것이라고 분석되었다. 따라서 유사성은 두 문장간의 일치하는 단어의 수에 따라 유사 또는 비유사로 구분할 수 있다.

이러한 지식을 바탕으로 하여 문장간의 유사도 측정 공식을 유도하였는데 절간에는 섀튼의 유사도 측정식을, 문장간에는 코싸인계수를 사용하였다. 절간의 유사성을 조사하여 유사하다고 판정받은 문장들을 그의 유사도

측정식에 대입한 결과 한계치가 약 0.1이었다. 그리고 문장의 경우를 코싸인계수로 산출하였더니 그 한계치는 0.66이었다.

유사문장들의 통합은 두가지 과정을 거쳐 수행하였는데 첫째가 여러 문장 중에서 대표될 만한 문장을 뽑아 그 문장으로 하여금 여러문장을 대변토록 하는 ‘문장 대표 선정’ 방법이었다. 둘째로는 여러 문장들을 분해하여 그 분해된 부분으로부터 대표문장을 생성하여 내는 ‘문장 대표 생성’ 방법을 적용시켰다. 후자의 통합과정에서 절과 문장의 템프릿을 작성하였고 또 5개의 연산자(통합가능연산자, 절선택연산자, 절조합연산자 A와 B와 C형)를 가동하였다.

본 연구에서 추구한 바는 자동초록을 수작업 수준으로 올리는 데의 기초가 될 수 있는 지식을 생산하는 것이었다. 앞으로의 연구에서 ‘절별 대체’는 보다 자유롭게 문장을 만들

참 고 문 헌

- 수 있는 '단어 대체'로 전환되어야 할 것으로 생각한다.
- 윤평현. 1992. 『국어의 접속어미 연구-의미론적 기능을 중심으로-』. 서울: 한신문화사.
- 이삼형. 1994. 『설명적 텍스트의 내용 구조 분석방법과 교육적 적용 연구』. 서울대학교 대학원 박사학위논문.
- 이상태. 1995. 『국어이음월의 통사·의미론적 연구』. 서울: 형실출판사.
- 이태영. 1998. 문장의 복합명사와 명사구의 유사정도에 대한 고찰. 『제6회 한국정보관리 학회 학술대회 논문집』, 43-46.
- Endres-Niggemeyer, B., A. Sigel, and E. Maier. 1995. "How to implement a naturalistic model of abstracting: four core working steps of an expert abstractor." *Information Processing & Management*, 31: 631-74.
- Karov, Y., and S. Edelman. 1998. "Similarity-based Word Sense Disambiguation.", *Computational Linguistics*, 24: 41-59
- McKeown, K., K. Kukich, and J. Robin. 1995. "Generating concise natural language summaries.", *Information Processing & Management*, 31: 703-33.
- Moens, M-F., C. Uyttendaele, and J. Dumortier. 1999. "Abstracting of Legal Cases: The Pontential of Clustering Based on the Selection of Representative Objects." *Journal of American Society for Information Science*, 50: 151-161.
- Radev, D.R. and K. R. Mckeown. 1998. "Generating Natural Language Summaries from Mutiple On-line Sources." *Computational Linguistics*, 24: 469-500
- Salton, G., J. Allen, and A. Singhal. 1996. "Automatic text decomposition and structuring." *Information Processing & Management*, 32: 127-138.
- Salton, G., Singhal, A, Mitra, M, Buckley, C, 1997. "Automatic text structuring and summarization." *Information Processing & Management*, 33: 193-207.

Schutze, H. 1998. "Automatic word sense discrimination." *Computational Linguistics*, 24: 97-123.

<부 록>

다음의 (1)에서 (27)까지 이어지는 질문들은 두 문장의 유사 여부를 묻는 것입니다. '㉑:㉒'는 알파벳 원 문자 ㉑와 ㉒로 구별된 문장들이 서로 유사한가를 묻는 표식입니다. 유사 정도는 100%, 75%, 50%, 25%, 0%로 나타냅니다.(50%가 조금 넘는다고 생각하면 50+%로 표시합니다.)

질문(1) : 예1)에서 ㉑:㉒=(100%), ㉑:㉓=(100%), ㉒:㉓=(100%)의 유사정도를 옆의 괄호 안에 적어 주시오

예1)

㉑ 열람목록은 이용자에게 도서관의 장서 전체를 보여 준다

㉒ 이용자는 열람목록으로 도서관의 장서 전체를 볼 수 있다

㉓ 도서관의 장서 전체는 열람목록에서 이용자한테(로) 보여진다

질문(2) : 예2)에서 ㉑:㉒=(50+%), ㉑:㉓=(75%), ㉑:㉔=(100%), ㉒:㉓=(50+%)의 유사정도를 옆의 괄호 안에 적어 주시오?

예2)

㉑ 도서관학은 문헌정보학에 포함되었다

㉒ 문헌정보학은 도서관학을 포함할 수도 있다

㉓ 도서관학이 문헌정보학으로 포함되어질 것이다

㉔ 문헌정보학이 도서관학을 포함한다

질문(3) : 예3)에서 ㉑:㉒?(50+%), ㉑:㉓?(50+%), ㉒:㉓?(75%)의 유사정도를 옆의 괄호 안에 적어 주시오

예3)

㉑ 참고봉사는 정보학, 서지학, 전산언어학, 언어통계학을 응용한다

㉒ 자동목록은 정보학, 서지학, 전산언어학, 언어통계학을 응용한다

㉓ 자동색인은 정보학, 서지학, 전산언어학, 언어통계학을 응용한다

질문(4) : 예4)에서 ㉑:㉒?(50+%), ㉑:㉔?(50+%), ㉓:㉔?(75%)의 유사정도를 옆의 괄호 안에 적어 주시오

예4)

㉑ 참고봉사, 자동목록, 자동색인은 서지학을 필요로 한다

㉓ 참고봉사, 자동목록, 자동색인은 전산언어학을 필요로 한다

㉔ 참고봉사, 자동목록, 자동색인은 언어통계학을 필요로 한다

질문(5) : 예5)에서 ㉑:㉒?(50%), ㉑:㉓?(50%), ㉒:㉓?(75%)의 유사정도를 옆의 괄호 안에 적어 주시오

예5)

- ㉔참고봉사는 서지학을 응용한다
- ㉕자동목록은 서지학을 응용한다
- ㉖자동색인은 서지학을 응용한다

질문(6) : 예6)에서 ㉔:㉕?(50%), ㉔:㉖?(50%), ㉕:㉖?(75%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예6)

- ㉔자동색인은 서지학을 필요로 한다
- ㉕자동색인은 전산언어학을 필요로 한다
- ㉖자동색인은 언어통계학을 필요로 한다

질문(7) : 예7)에서 ㉔:㉕?(0%), ㉔:㉖?(0%), ㉕:㉖?(50+%), ㉗:㉘?(0%), ㉗:㉙?(0%), ㉘:㉙?(25%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예7)

- ㉔참고봉사는 응용된다
- ㉕자동목록은 응용된다
- ㉖자동색인은 응용된다.
- ㉗서지학을 필요로 한다
- ㉘전산언어학을 필요로 한다
- ㉙언어통계학을 필요로 한다

질문(8) : 예8)에서 ㉔:㉕?(0%), ㉕:㉖?(50+%), ㉕:㉗?(0%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예8)

- ㉔정치학을 인용한다
- ㉕경제학을 인용한다
- ㉖경제학을 다룬다
- ㉗경제학을 무시한다.

질문(9) : 예9)에서 ㉔:㉕?(0%), ㉔:㉖?(25%), ㉔:㉗?(0%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예9)

- ㉔정치학을 경제학 문헌이 인용한다
- ㉕물리학이 경제학 문헌에 인용된다
- ㉖정치학을 경제학 문헌이 정리한다
- ㉗정치학을 경제학 문헌이 배경한다

질문(10) : 예10)에서 ㉔:㉕?(75%), ㉔:㉖?(50%), ㉕:㉖?(50+%), ㉕:㉗?(75%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예10)

- ㉠사회와 문화의 발전을 가져왔다
- ㉡사회와 문화의 발전을 촉진한다
- ㉢사회와 문화의 발전을 홍보한다
- ㉣사회와 정치의 발전을 촉진한다

질문(11) : 예11)에서 ㉠:㉡?(50%), ㉡:㉢?(50%), ㉡:㉣?(25%), ㉠:㉣?(75%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예11)

- ㉠:㉡ 사회의 발전을 촉진한다
- ㉡:㉢ 정치의 발전을 촉진한다
- ㉡:㉣ 사회를 가져온다
- ㉠:㉣ 정치를 촉진한다
- ㉠:㉣ 사회를 촉진한다
- ㉡:㉣ 정치를 가져왔다

질문(12) : 예12)에서 ㉠:㉡?(50%), ㉠:㉢?(50%), ㉡:㉢?(50%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예12)

- ㉠:㉡ 발전을 가져온다
- ㉡:㉢ 발전을 촉진한다
- ㉡:㉢ 발전을 홍보한다

질문(13) : 예13)에서 ㉠:㉡?(100%), ㉠:㉢?(75%), ㉠:㉣?(100%), ㉡:㉣?(50+%), ㉢:㉣?(75%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예13)

- ㉠:㉡ 온라인 목록은 이용자가 많은 접근점을 갖게 되어 도서내용을 잘 탐색할 수 있다.
- ㉡:㉢ 온라인 목록은 도서내용을 잘 탐색할 수 있게 많은 접근점을 갖는다.
- ㉢:㉣ 온라인 목록은 이용자가 많은 접근점을 갖게 되면 도서내용을 잘 탐색할 수 있다.
- ㉣:㉣ 온라인 목록은 이용자가 많은 접근점을 갖게 되며 따라서 도서내용을 잘 탐색할 수 있다.

질문(14) : 예14)에서 ㉠:㉡?(75%), ㉠:㉢?(75%), ㉠:㉣?(75%), ㉠:㉣?(50+%), ㉠:㉣:(75%), ㉠:㉣:(50+%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예14)

- ㉠:㉡ 역시 사회학은 사회과학적 성향이 강하며 경제학 및 정치학 분야 문헌을 선호하고 있음을 알 수 있다.
- ㉠:㉢ 역시 사회학은 사회과학적 성향이 강하므로 경제학 및 정치학 분야 문헌을 선호하고 있음을 알 수 있다.
- ㉠:㉣ 역시 사회학은 사회과학적 성향이 강해서 경제학 및 정치학 분야 문헌을 선호하고 있음을 알 수 있다.

알 수 있다.

㉠역시 사회학은 사회과학적 성향이 강하여 경제학 및 정치학 분야 문헌을 선호하고 있음을 알 수 있다.

㉡역시 사회학은 사회과학적 성향이 강한데 경제학 및 정치학 분야 문헌을 선호하고 있음을 알 수 있다.

㉢역시 사회학은 사회과학적 성향이 강하면 경제학 및 정치학 분야 문헌을 선호하고 있음을 알 수 있다.

㉣사회학은 경제학 및 정치학 분야 문헌을 선호하는데 이는 사회과학적 성향이 강하기 때문이다

질문(15) : 예15)에서 ㉠:㉡?(50%), ㉠:㉢?(25%), ㉡:㉢?(25%)의 유사정도를 옆의 괄호 안에 적어 주시오
예15)

㉠사회학은 경제학 및 정치학 분야 문헌을 선호하는데 정치학은 철학이나 경영학에서 많은 영향을 받는다.

㉡사회학은 경제학 및 정치학 분야 문헌을 선호하면 정치학은 철학이나 경영학에서 많은 영향을 받는다.

㉢사회학은 경제학 및 정치학 분야 문헌을 선호하는데 이는 사회과학적 성향이 강하기 때문이다

질문(16) : 예16)에서 ㉠:㉡?(50+%), ㉠:㉢?(75%), ㉡:㉢?(100%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예16)

㉠책자형 목록은 카드목록과 함께 사용되기는 하지만 도서관의 자료검색을 위해서는 속보성이 결여되어 있다.

㉡속보성이 결여되어 있는 책자형 목록은 도서관의 자료검색에 문제점을 제기한다.

㉢책자형 목록은 속보성이 결여되어 있어 도서관의 자료검색에 문제점을 제기한다.

질문(17) : 예17)에서 ㉠:㉡?(25%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예17)

㉠주제를 규명하였다

㉡정보전달을 규명하였다

질문(18) : 예18)에서 ㉠:㉡?(50%), ㉠:㉢?(0%), ㉡:㉢?(25%), ㉠:㉣?(50+%), ㉠:㉤?(50%), ㉡:㉥?(50+%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예18)

㉠사서가 주제를 선정하였다

㉡사서가 책을 선정하였다

㉢이용자가 책을 선정하였다

㉣사서가 주제를 연구하였다

㉔주제를 선정하였다

㉕사서가 선정하였다

질문(19) : 예19)에서 ㉔:㉕?(50+%), ㉔:㉖?(50+%), ㉔㉕?(25%), ㉕:㉖?(25%), ㉕:㉕?(25%), ㉔:㉕:㉖?(50+%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예19)

㉔열람목록은 이용자에게 도서관 소장물을 보여준다

㉕색인은 이용자에게 도서관 소장물을 보여준다

㉖열람목록은 사서에 대해 도서관 소장물을 보여준다

㉗열람목록으로 사서가 도서관 소장물을 정리한다(처리한다)

질문(20) : 예20)에서 ㉔:㉕?(75%), ㉔:㉖?(75%), ㉕:㉖?(50+%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예20)

㉔카드목록은 다른 보완도구와는 별도로 그 내용이나 체재기능면에서 특별한 도구로서 발전되어 왔다

㉕카드색인은 다른 보완도구와는 별도로 그 내용이나 체재기능면에서 특별한 도구로서 발전되어 왔다

㉖책자목록은 다른 보완도구와는 별도로 그 내용이나 체재기능면에서 특별한 도구로서 발전되어 왔다

질문(21) : 예21)에서 ㉔:㉕?(0%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예21)

㉔간략소개목록으로서 필요한 최소량의 기록을 갖고 매 기록당 한정된 접근점을 허용하여 풍부한 접근과 탐색능력을 갖추어야 할 것이다

㉕다른 형태의 목록에서 가능한 것보다 많은 접근점, 기능성, 이용용이성을 가진 목록으로 규정하고 있다

질문(22) : 예22)에서 ㉔:㉕?(75%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예22)

㉔안내스크린은 코멘드의 예와 그 시스템에 대한 간단한 정의와 범위를 보여주어야 한다

㉕Carlyle 시스템의 안내스크린은 그 시스템에 대한 설명과 사용법, 사용되는 코멘드의 실례, 도움을 더 얻을 수 있는 방법을 보여준다

질문(23) : 예23)에서 ㉔:㉕?(50+%), ㉔:㉖?(50%), ㉕:㉖?(50+%), ㉔:㉕:㉖?(50%), ㉔:㉕:㉖:㉗?(75%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예23)

㉔타 분야 문헌의 주제별 분포를 대체적으로 규명하였다

㉕타 분야 문헌들의 주제별 분포를 중점적으로 규명하였다

㉔그 문헌들의 주제별 분포를 중점적으로 규명하였다

㉕그 것들의 주제별 분포를 중점적으로 규명하였다

질문(24) : 예24)에서 ㉔:㉕?(50+%), ㉔:㉖?(100%), ㉕:㉖?(75+%), ㉔:㉗?(50%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예24)

㉔그 것들의 주제별 분포를 중점적으로 규명하였다

㉕그 것들의 주제별 분포를 역시 규명하였다

㉖그 것들의 주제별 분포를 역시 중점적으로 규명하였다

㉗그 것들의 주제별 분포를 첫째로 규명하였다

㉘그 것들의 주제별 분포를 셋째로 규명하였다

질문(25) : 예25)에서 ㉔:㉕?(0%), ㉔:㉖?(50%), ㉕:㉖?(0%), ㉔:㉗?(75%), ㉕:㉘?(25%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예25)

㉔정치학은 역시 사회과학적 성향이 높다는 것을 이 표에 의해 유추할 수 있다

㉕경제학은 역시 사회과학적 성향이 높다는 것을 이 표에 의해 유추할 수 있다

㉖정치학은 역시 법학과 관련이 높다는 것을 이 표에 의해 유추할 수 있다

㉗정치학은 역시 사회과학적 성향이 높다는 것을 그 표는 말해 주고 있다

질문(26) : 예26)에서 ㉔:㉕?(50%)의 유사정도를 옆의 괄호 안에 적어 주시오.

예26)

㉔이 연구의 목적을 달성하기 위하여 타분야 문헌의 주제별 분포, 정보전달 매체별 분포 및 국가별 분포를 구명하였다

㉕선정된 학문분야의 문헌에 인용된 타분야 문헌들의 주제별 분포, 정보전달매체별 분포 및 국가별 분포를 중점적으로 구명하였다

* 다음의 '~ ~'로 구분되어 있는 문장집단들을 읽고 각 집단의 문장들이 비슷한 것끼리 짝지어 있는지(유사정도가 50% 위) 아니면 비슷하지 않은 것이 포함되어 있는지(유사정도가 50% 아래)를 구별합니다. 비슷하다고 생각하는 문장들의 알파벳 원문자를 문장집단 밑의 여백에 각각 모아서 적어 주시오

질문(32) : 제목: 사회학과 경제학의 의존도 비교(제목을 참조하여 판단합니다)

㉔표 1에 제시된 바와 같이 우리나라 경제학 분야의 타분야 문헌의존도는 6.30퍼센트이다.

㉕표 1에 제시된 바와 같이 우리나라 사회학 분야의 자분야 의존도는 80.34퍼센트이고 타분야 의존도는 19.66퍼센트이다.

㉖우리나라 정치학 문헌은 표 1에 제시된 바와 같이 자분야 의존도 85.64퍼센트와 타분야 의존

도 14.36퍼센트를 기록하고 있다.

a-b

질문(33) :

- ㉠그렇게 구명된 연구결과는 우리나라 학술 정보시스템의 효율적 운영에 기여할 수 있을 것이다.
- ㉡전술한 몇가지 결론은 우리나라 학술정보시스템의 효율적 운영을 돕는 기초자료가 될 것이다.
- ㉢그렇게 구명된 연구결과는 일본의 학술 정보시스템의 효율적 운영에 기여할 수 있을 것이다.
- ㉣그렇게 구명된 연구결과는 우리나라 정치시스템의 효율적 운영에 기여할 수 있을 것이다.
- ㉤전술한 몇가지 결론은 우리나라 정치시스템의 효율적 운영을 돕는 기초자료가 될 것이다.
- ㉥결론적으로 보아 우리나라 학술 정보시스템의 발전적 운영에 혁신을 이루었다고 할 수 있을 것이다.
- ㉦결론적으로 보아 미국 학술 정보시스템의 발전적 운영에 혁신을 이루었다고 할 수 있을 것이다.

a-b, a-c, a-f, d-e, c-d

질문(34)

- ㉠㉡온라인 열람목록은 (㉠지역사회 주민 및 도서관 이용자의 복리와 검색 효율성을 촉진시키고) (㉡(정보를 획득하는) 작업을 끊임없이 향상시킬 수 있고), (㉢(필요한 때는 시간과 장소에 구애받지 않고) (정보를 획득할 수 있어야 하며)), (㉣(다른 형태의 목록과 비교하여) (유지비가 경감되게 고안되어야 한다.))
- ㉠㉢온라인 목록은 (㉠저자, 서명, 주제어 등의 접근을 통하여) (㉡목록정보를 빠르고 정확하게 검색할 수 있도록 하고) (㉢주거지, 직장, 교육기관과 같은 다양한 장소에서 정보획득이 가능하고) (㉣다양한 자료에 접근과 유지가 쉬어야 할 것이다.)

a-b

질문(35)

- ㉠㉡온라인 열람목록은 (㉠지역사회 주민 및 도서관 이용자의 복리와 검색 효율성을 촉진시키고) (㉡(정보를 획득하는) 작업을 끊임없이 향상시킬 수 있고), (㉢(필요한 때는 시간과 장소에 구애받지 않고) (정보를 획득할 수 있어야 하며)), (㉣(다른 형태의 목록과 비교하여) (다양한 자료에 쉽게 접할 수 있어야 한다.))
- ㉠㉢온라인 목록은 (㉠저자, 서명, 주제어 등의 접근을 통하여) (㉡목록기입을 빠르고 정확하게 검색할 수 있도록 하고) (㉢주거지, 직장, 교육기관과 같은 다양한 장소에서 정보획득이 가능하고) (㉣지역적, 국가적, 국제적으로 다양한 자료에 접근이 가능하도록 하여야 할 것이다)
- ㉠㉣온라인 목록은 (㉠필요한 때에 어디서나 정보를 검색할 수 있게) (㉡네트워크가 구축되어야 한다)

a-b