분야연상어를 이용한 화제의 계속성과 전환성을 추적하는 단락분할 방법

이 상 곤†

요 약

복수의 화제가 혼합되어 있는 문서에서 각 화제의 경계부분을 구분하여 결정하는 기술을 단락분할이라 한다. 이 기술은 정보검색의 분야에만 한정되지 않고 다양한 분야에서 중요한 역할을 담당할 기술이다. 잘 정의된 분야체계에 따라 구축된 분야연상어를 이용하여 단락분할을 시도한다. 분야연상어란 특정한 분야를 정확하게 연상할 수 있는 단어로서 잘 분류된 문서 컬렉션에서 구축할 수 있다. 이 분야연상어를 이용하여 문서를 관련된 분야별로 추출하여 의미기반 단락추출 방법을 제안한다. 화제의 계속성에 주목하여 분야연상어의 수준(범위)이나 연속출현성에 의해 계산된 계속도에 의해 화제의 실마리를 추적하고, 화제의 전환성을 고려한 방법을 제안한다. 문서 내 각 화제의 단락구분을 명확히 하여, 단락을 화제분야별로 추출하는 방법을 제안한다. 일본어 50문서를 실험한 결과 82%의 정확율과 63%의 재현율을 얻어 실용성을 기대할 수 있었고, 한국어에 적용하여도 좋을 것으로 예상한다.

Passage Retrieval based on Tracing Topic Continuity and Transition by Using Field-Associated Term

Samuel Sangkon Lee†

ABSTRACT

We propose a technique to extract a relevant passage from text collection based on field-associated terms since they tries to concentrate relevant text to users query. Documents are supposed to be managed as a whole without any segmentation into small pieces, but the method presented is independent upon any text-embedded auxiliary information, and is based on topic continuity and transition. For users needs—relative sentences or passages, We present a passage retrieval techniques by using occurrence frequency of a field-associated term to delimit text, that is likely to be relevant to a particular topic, considering continuity and transition within topic flowing in text. We evaluate 50 Japanese documents and verify the usefulness with 82% for average precision and 63% for recall.

키워드 : 분야연상어(Field-Associated Term), 화제 계속성과 전환성, 단락분할(Passage Retrieval), 문서분류, 정보검색

1. 서 론

복수의 화제가 혼합되어 있는 문서에서 화제의 실마리 부분을 특정화하여 각 화제별 단락을 추출하는 기술은 정보검색분야에서 중요한 역할을 담당하는 기술이다. 대량의 전자문서에서 사용자의 검색요구에 맞는 문서를 검색하는 종래의 검색방법은 문서전체를 하나의 객체로 생각하여 검색요구에 적당한 문서를 검색하여 왔다. 그러나 실제문서에는 복수의 화제가 혼합되어 있기 때문에 문서전체를 검색대상으로 하지 않고, 검색요구에 정확히 일치하는 텍스트 단편만을 검색하는 단락검색(PR ; Passage Retrieval[2-5, 9-15]) 기술이 미국이나 일본 등지에서 주목을 받고 있으나, 한국에서의 연구는 활발하지 못한 실정이다.

대량의 문서를 특정한 기준에 따라 자동으로 분류하는 연구분야는 각 단락간의 유사도를 계산하여 유사도가 높은 순으로 문서를 분류하는 단락별 분류방법[11, 19-21]이 사용되고 있다. 또한 문서에서 특정한 정보를 추출하는 문서추출 분야에서도 문서의 화제분야 추정은 유용한 실마리가 된다.

이상의 연구분야 이외에도 넓은 의미의 단락검색 기술은 유용하다. 예를 들면, 음성대화 시스템에서 대화가 진행되고 있는 화제분야가 정해지면 그 분야에 해당하는 대화모

† 정 회 원 : 전주대학교 정보기술컴퓨터공학부 교수
논문접수 : 2002년 7월 22일, 심사완료 : 2003년 1월 3일

벨의 추정이 가능하게 된다. 일본어 문서 처리의 경우 가명한자 변환시스템(Kana-to-Kanji Conversion System)[24]에서 화제분야를 이용하여 동음이의어 등의 가능한 변환후보를 추출한다. 문서내 도표와 도표의 설명부분(단락에 해당)을 추출하는 기술[20]도 문서를 구조화하여 상세히 구축할 수 있다. 이와 같이 단락검색은 문서 내에 존재하는 각각의 화제부분을 추출하여 추출 정밀도를 크게 향상시킬 수 있으며, 정보검색, 문서분류, 정보추출 등 다양한 연구분야에 적용할 수 있는 기술이다.

종래의 미국이나 일본의 연구는 키워드로 일반적인 단어를 이용하고 있으며, 각 단어가 분야를 한정하는 정도나 범위를 세밀하게 고려하고 있지 않기 때문에 검색의 정밀도가 떨어지는 경향이 있다.

한편, Takako[23]는 미리 정의된 분야체계에 따라 문서 데이터에서 각 분야 특유의 분야연상어를 구축하는 방법을 제안하고 있다. 분야연상어란 인간이 문서에서 '투수'나 '선거'와 같은 단어를 보는 것만으로 <야구>[1]나 <정치>라고 하는 상식적 분야를 인지할 수 있는 단일어 또는 복합어로 된 상식적인 단어를 말한다.

본 논문에서는 분야연상어를 이용하여 검색요구에 일치하는 단락의 추출을 목적으로 한다. 단락이 문서의 특정화제에 대하여 쓰여진 것인지를 판별하기 위해 단락의 범위를 결정하고, 그 단락의 분야를 결정하는 방법을 제안한다. 분야연상어가 나타나는 텍스트 주변부분은 특정화제를 묘사하는 화제분야로 추정할 수 있다. 그러나, 분야연상어가 나타나지 않는 단락에 대해서는 어떻게 분야를 결정할 것인가의 문제가 있다. 문서 내 화제를 파악하고, 분야연상어의 연속 출현율을 기반으로 화제의 계속도를 계산하는 방법을 소개한다. 이와 동시에, 화제의 전환성을 고려하여 인접한 문장간의 구간분리를 명확히 하는 방법을 소개한다. 이러한 방법을 이용하여 분야의 중복이 없는 단락을 추출한다.

이하 제 2장에서는 본 논문과 동일한 연구과제에 대한 종래의 연구에 대하여 살펴보고, 제 3장에서는 미리 정의된 분야체계에 따라 각 분야를 지시하는 분야연상어에 대하여 설명한다. 제 4장에서는 일반적인 문서에서 화제흐름의 특징을 정의하고, 문서 내에서 동일분야의 단락을 결정하는 방법을 설명한다. 제 5장에서는 예제 문서를 사용하여 실험을 하고, 본 방법의 유용성을 평가한다. 마지막 장에서는 결론과 향후의 과제에 대하여 서술한다.

1) 분야연상어('과'을 사용)와 구분하기 위해 분야명을 기호 '<' 과 '>' 내에 기술한다.

2. 종래의 연구

현재까지의 단락검색에 관한 연구는 검색단위로 문서내의 장, 절, 혹은 단락과 같은 형식적인 정보에 기반을 둔 연구와 고정길이 혹은 가변길이의 윈도우에 기초를 둔 연구, 그리고 문서의 형식적인 정보에 의존하지 않고 의미적 실마리에 기반을 둔 연구 등 크게 세종류로 나눌 수 있다.

문서의 형식적인 구조에 기반한 방법은 대략적인 단락의 추출은 비교적 쉽게 할 수 있다. 그러나, 단락이나 장, 절에 복수의 화제가 존재하거나, 복수의 화제가 여러 단락이나 장, 절에 걸쳐 기술되어 있는 경우에는 문서의 형식적인 정보만으로는 단락검색의 결과가 상세하고 유연하게 결정되지 못하는 문제가 있다.

고정길이 윈도우에 의한 방법[3]은 각 문서마다 윈도우를 슬라이드 시켜 검색요구와 가장 관련이 깊은 단락으로 결정한다. 가변길이 윈도우에 의한 방법은 고정길이 윈도우에 의한 방법과 동일하게 윈도우를 단락단위로 검색하거나 일정한 범위 내에서 윈도우의 폭을 변화시켜 단락의 범위를 조정한다. 그러나, 단락의 의미적인 실마리를 형성하는 곳의 유연성이 좋지 못하며 처리비용도 높아 실용성 면에서 양 방법 모두 유용한 방법이라고 말할 수 없다.

의미적인 실마리에 기초한 방법으로 Hearst 등[8]은 사전에 문서를 고정길이의 블록으로 분할하고 블록 내에 출현하는 단어의 결속성과 유사한 블록을 실마리로 하여 단락을 형성하는 방법을 제안하고 있으나, 문서에서 표층적으로 나타나는 단어정보만을 이용하고 있다는 점에서 개선의 여지가 있다. Mochizuki 등[21]은 검색요구의 단어와 그 단어와 동일개념을 갖는 단어 또는 공기관계가 있는 단어가 출현하는 위치에서 동일개념의 어휘적인 연쇄를 계산하여 단락을 추출하는 방법을 제안하였다. 어휘적 연쇄가 구해지면 각 단어에 대한 연쇄의 길이나 갭(gap) 길이는 일정한 값으로 설정된다. 그러나 연쇄를 형성하는 단어들 간의 연속출현성과 출현분포는 고려하지 않고 있다.

Mizuno 등[20]은 단락검색의 응용분야로서 문서의 각 도표에 대한 설명부분의 범위를 한정하는 방법을 제안하였다. 이 방법에서는 도표중이나 캡션 내에서 사용되는 단어를 키워드로 하여, 문서 내에서 키워드의 출현밀도 분포가 높은 부분을 그 도표의 설명부분으로 간주한다. 여기서 각 키워드의 출현율이 집중되는 정도를 표현하는 편차의 산출에는 Harming 윈도우 함수로 계산되는 출현밀도를 사용하고 있다. 이 방법은 중요한 키워드의 출현분포에 기반하기 때문에 다소 정밀하게 단락을 형성할 수 있으나, 단락의 경계부분에서 정밀도에 문제[2]가 있다. 또한, 설명내용의 화제가

전환되는 성질을 고려하고 있지 않으며, 각 설명구간의 분할을 명확하게 하는 방법에 대해서도 언급하고 있지 않다. 그 때문에 문서내 특정 부분에서 복수의 도표가 밀접하고 각 도표의 설명부분이 인접하고 있는 경우는 설명부분이 중복할 염려가 있다.

본 논문의 방법은 Mochizuki 등[21]과 같이 문서 내에 출현하는 모든 단어정보를 사용하여 분야 한정에 유일한 분야의 연상정보를 사용하여 문서의 의미적 실마리에 유연하게 단락을 한정한다. 그러나, [21]과 다른 점은 화제의 계속성을 고려하여 동일한 분야를 연상하는 분야연상어가 연속하여 출현하는 경우를 분포정보로 도입한다. 또한, 화제의 전환성을 고려하여 복수의 분야에 대한 단락이 연속하고 있는 경우에도 단락간의 중복을 피하여 각 단락의 구간분리를 명확하게 하는 유연성 있는 방법을 제시한다.

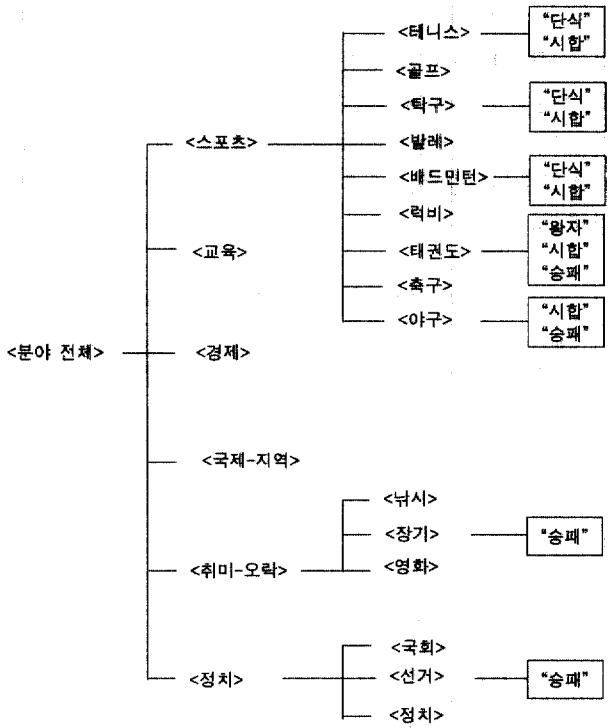
3. 분야연상어

3.1 분야 체계

분야 체계란 각 분야의 상위·하위관계를 트리구조로 표현한 분야별 체계를 말한다. 이를 "분야트리"라 정의하고, 분야트리의 잎에 상응하는 분야를 "종단분야", 종단분야 이외는 "중간분야"라 부른다. 본 연구는 일본어 용어집 Imidas[6]를 표본으로 분야트리를 구축하였다. 이 분야트리의 전체분야 수는 443개이며, 중간분야 수는 50개, 종단분야는 393개(깊이 2, 3, 4의 종단분야는 각각 174개, 208개, 11개)이다. 직접 상위분야 혹은 하위분야를 각각 "부모분야", "자식분야"라 부른다. 분야의 지정은 분야명의 패스 <P>로 기술하지만, 뿌리에 상응하는 <전체분야>는 생략하여 기술하는 것을 원칙으로 한다. 특히 모순이 생기지 않는 경우는 전체 패스명을 생략하고, 종단분야만으로 설명한다. (그림 1)에 분야트리의 예를 표시하였다. 예를 들어, 어떤 분야의 패스 <P>가 <스포츠/배드민턴>이면 <스포츠>의 하위분야 <배드민턴>만으로 표시한다.

분야트리에 따라 미리 분류한 문서 데이터(문서 컬렉션)를 형태소 해석하고, 각 문서 내에 존재하는 분야연상어를 추출한다. 각 분야에 속하는 문서 데이터 내에 출현하는 분야연상어의 집중율을 계산한다. 여기서 구해진 분야연상어는 형태소사전에 등록되어 있는 단어(단일어 혹은 단위어)이며, 복합어에 대한 분야연상어는 단일어의 분야계승에 기초하여 반자동적으로 구축할 수 있다. 형태소 해석 결과, 미등록어가 되는 단어는 분야연상어의 대상으로 하지 않는다. (그림 1)에 분야트리와 함께 분야연상어의 예도 표시하였다.

2) 재현율은 높지만, 정확율은 낮다.



(그림 1) 분야트리와 분야연상어의 예

3.2 분야연상어의 수준

문서 컬렉션에서 추출한 분야연상어에는 연상되는 분야의 넓이에 차이가 있다. 이 단어는 유일한 종단분야나 중간분야를 지시하는 단어 혹은 복수의 종단분야나 중간분야를 지시하므로, 각 분야연상어 w의 수준을 아래와 같이 정의한다[23].

(수준 1) 완전연상어 : w는 유일의 종단분야만을 연상한다.
(수준 2) 준완전연상어 : w는 같은 부모분야를 갖는 종단분야 중에서 한정된 복수의 종단분야를 연상한다.
(수준 3) 중간연상어 : w는 완전연상어, 준완전연상어가 아니고, 유일의 중간분야만을 연상한다.
(수준 4) 다분야연상어 : w는 완전연상어, 준완전연상어, 중간연상어가 아니고, 복수의 중간분야와 종단분야를 연상한다.
(수준 5) 비연상어 : w는 위의 수준 1~4 이외이고, 어떠한 특정분야도 연상하지 않는다.

(그림 1)에서 예시한 분야트리에 따라 각 분야연상어의 수준을 위의 <표 1>에 표시하였다.

수준 1의 완전연상어에서 '국기원'은 종단분야 <태권도>를 오직 하나의 분야로 한정한다. 수준 2의 '단식'과 '복식'은 준완전연상어인데, 부모분야 <스포츠>내에서 복수의 종단분야 <테니스>, <탁구> 혹은 <배드민턴> 등을 한정한다. 수준 3의 중간연상어 '시합'은 어떠한 종단분야도 한정

하지 않으나, 한 개의 중간분야 <스포츠>를 한정한다. 또한, 수준 4의 다분야연상어 '승패'는 중간분야 <스포츠> 혹은 복수의 종단분야 <취미·오락/장기>, <정치/선거> 등 복수의 분야를 한정할 수 있는 분야연상어이다. 마지막으로 수준 5의 비연상어는 '경우', '사용'과 같이 어떤 특정분야도 한정하지 않는 단어이다.

〈표 1〉 각 수준별 분야연상어의 예

연 상 어	연 상 분 야	수 준
국기원	<스포츠/태권도>	1
단식, 복식	<스포츠/테니스> <스포츠/탁구> <스포츠/배드민턴>	2
시 합	<스포츠>	3
승 패	<스포츠> <취미·오락/장기> <정치/선거>	4
경우, 사용	-	5

4. 단락의 결정

본 방법에서는 문서의 각 문장마다 처리를 진행해 분야별 단락을 추출한다. 이하 설명에서 사용되는 각각의 변수를 정의한다. 먼저, 처리대상 문서 $d_i = \{s_{i,1}, s_{i,2}, s_{i,3}, \cdots, s_{i,j}, \cdots, s_{i,m}\}$이다. 단, $s_{i,j}$는 문서 d_i 내의 j번째 문을 표시한다. F는 분야트리 전체집합을 의미하며, $\{F_1, F_2, F_3, \cdots, F_k, \cdots, F_n\}$으로 구성되어 있다. $Frequency(s_{i,j}, F_k)$는 문서 d_i의 한 문장 s_j 내에 존재하는 분야 F_k의 분야연상어의 점수(분야연상어의 세기를 의미)이다. $Passage(F_k) = \{P_{k,1}, P_{k,2}, \cdots, P_{k,p}, \cdots\}$는 문서 d_i 내에 존재하는 분야 F_k의 단락의 집합으로 정의한다. 단, $P_{k,p}$은 문서 d_i 내에 존재하는 분야 F_k의 p번째 단락집합을 표시한다.

4.1 분야연상어의 점수집계

본 방법에서는 문서 내에 존재하는 분야연상어를 각 문장에서 추출한다. 추출할 때 복수 키워드에 대한 고속 문자열 조합법으로 알려진 AC법[1]을 이용한다. 미리 인간이 자신의 상식지식으로 구축한 분야연상어를 AC법을 이용해 AC 사전으로 저장하여 두고, 각 문장에 존재하는 모든 분야연상어를 추출한다. 추출된 분야연상어는 각 수준에 따라 분야를 한정하는 정도가 다르기 때문에 동일분야에 대한 각 수준별 점수를 합산한다. 각 수준의 점수로서 수준 1을 10, 수준 2를 5, 수준 3을 3, 수준 4를 2점으로 각각 설정한다.

아래의 예제문장(각 문장의 $s_{i,1} \sim s_{i,4}$가 <야구>, $s_{i,5} \sim s_{i,6}$이 <축구>의 화제에 대하여 쓰여진 문서)에 대하여 점수집계를 하면 이탤릭·볼드체 단어가 분야연상어로 추출된다. 여기서 각 분야연상어의 오른쪽 위에 기술한 위첨자는 분야연상어의 각 수준을 표시한다. 분야연상어 "시합"과 "선수"는 <스포츠>를, "홈런"은 <스포츠/야구>를 "축구"와 "MF"는 <스포츠/축구>를 연상한다. 이 결과를 기초로 점수 결과를 집계하여 보면 <표 2>와 같다.

〈표 2〉 예제 문장의 점수집계 결과

문장 번호	야 구	테니스	축 구
$s_{i,1}$	3	3	3
$s_{i,2}$	10	0	0
$s_{i,3}$	0	0	0
$s_{i,4}$	13	3	3
$s_{i,5}$	0	0	10
$s_{i,6}$	0	0	10

◎ 예제

오늘 체육시간에 ***시합***^{3}이 있었다. $s_{i,1}$

나는 ***선수***^{1}로 출전한다. $s_{i,2}$

비가 우려되었으나 아무 문제없었다. $s_{i,3}$

시합^{3}은 굿바이 ***홈런***^{1}으로 이겼다. $s_{i,4}$

다음 주에는 ***축구***^{1}가 있다. $s_{i,5}$

나는 ***MF***^{1}로 출전할 예정이다. $s_{i,6}$

각 문을 가장 점수가 높은 순으로 분야별 단락을 추출하면 Passage (<야구>) = $\{\{s_{i,1}, s_{i,2}\}, \{s_{i,4}\}\}$, Passage (<테니스>) = $\{\{s_{i,1}\}\}$[3], Passage (<축구>) = $\{\{s_{i,1}\}, \{s_{i,5}, s_{i,6}\}\}$의 단락이 형성된다. 그러나, 단락분할 결과 다음과 같은 두 가지 문제점이 있다.

① 분야연상어가 각 문장에서 계속하여 출현하지 않는 다면, 특정화제의 단락은 분리되어 화제의 실마리가 끊어진다.

예를 들면, <표 2>에서 $s_{i,3}$은 <야구>에 대한 분야연상어가 존재하지 않기 때문에 <야구>에 관한 Passage($s_{i,1} \sim s_{i,4}$)가 두 개의 작은 단락 ($\{s_{i,1}, s_{i,2}\}$과 $\{s_{i,4}\}$)으로 분리된다.

② 한 개의 문에서 복수분야의 분야연상어가 같은 점수로 출현하면 여러 분야에 속하는 단락이 형성되어 단락 들 사이에 분야별 중복이 발생한다.

3) 분야연상어 '시합'은 중간분야 <스포츠>를 한정하기 때문에 분야 <테니스>에 관한 분야연상어도 가능하다.

예를 들면, $s_{i,1}$이나 $s_{i,4}$는 복수의 분야(<야구>와 <테니스> 혹은 <야구>, <테니스>, <축구>)에 속한다. 이러한 문제들을 해결하기 위해 화제흐름의 특징을 고려하여 분야간 중복이 없는 단락을 추출하는 방법이 필요하다.

4.2 화제의 계속성과 전환성

신문이나 잡지기사 등의 일반적인 문서 내에 기술되는 화제의 흐름에는 다음의 두 가지 특징이 있다.

① 일련의 화제는 계속성을 가지며, 실마리를 형성하고 있기 때문에 한 개의 화제가 산발적으로 진행되는 일은 없다.

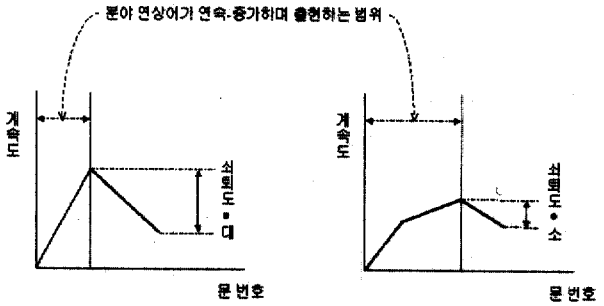
② 화제의 흐름에는 전환점이 있으며, 복수의 화제가 병행적으로 동시 진행되거나 중복되는 일은 없다.

위 ①의 특징에 의해 산발적으로 출현하고 있는 소규모의 단락을 화제의 "계속성"이란 척도로 정의한다. 화제의 계속성을 "계속도"라고 부르고, α로 명시한다. 또한, 특징 ②를 "전환성"이라는 관점에서 단락간 화제구간을 명확하게 한다. 복수분야의 화제로 양쪽 모두의 화제에 속하는 단락간 중복을 제거한다. 전환성을 측정하는 척도를 "전환도"라 정의하며, β로 표시한다. 다시 이들 특징을 이용하여 한 개의 문은 한 개 이하의 화제에 대응하기 때문에 처리대상의 문이 특정화제로 되는 분야를 "화제분야"라 정의하고 F_{theme}으로 표시한다. 이 화제분야 F_{theme}은 화제의 계속도에 의해 유지되고 화제의 전환도에 따라 변화한다.

화제분야 F_{theme}의 분야연상어가 연속해서 출현하면 계속도 α는 높아지고, 계속성이 끊어지면 계속도 α가 쇠퇴하는 것으로 설정된다. 또한, 전환도는 화제분야 F_{theme} 이외의 각 분야들에 준비되어 분야 F_k에 대한 전환도 $\beta(F_k)$는 $\alpha(F_k)$의 분야연상어가 계속하여 출현하는 정도를 나타내는 계속도와 비교하여 갱신된다. 다음에서는 계속도와 전환도의 구체적 계산방법에 대하여 고찰한다.

4.3 계속도와 전환도의 계산

본 방법에서는 계속도 α를 산출할 때에 화제의 계속성이 쇠퇴하는 비율을 "쇠퇴도"로 정의한다. (그림 2)(a)에서 표시한 바와 같이 텍스트의 좁은 범위에서 급격하게 출현하는 분야연상어를 포함되는 문은 화제의 계속성이 쇠퇴하기 쉽고, 거꾸로 (그림 2)(b)에서 표시한 바와 같이 광범위하게 끊임없이 출현하는 분야연상어를 포함하는 문장은 화제의 계속도가 높아져 화제성이 쇠퇴하기 어렵다는 성질을 이용한다. 여기서 분야연상어의 연속출현성을 고려한 문 s_i에서의 쇠퇴율(*Decline*)을 이하의 계산식으로 정의한다.



(a) 분야 연상어가 텍스트의 좁은 범위에서 급격히 출현하는 경우

(b) 분야 연상어가 텍스트의 넓은 범위에서 끊임없이 출현하는 경우

(그림 2) 화제의 연속 범위에 의한 쇠퇴도의 변화

$$Decline_{i,j} = \left[\frac{\sum_{s_{i,k} \in C_i} (Freq(s_{i,k}, F_{theme}))}{number(C_i) + 1} + \frac{Freq(s_{i,j}, F_{theme})}{number(C_i) + 1} \right]$$

단, $C_i = \{s_{i,j-n}, \cdots, s_{i,k}, \cdots, s_{i,j-1}\}$은 문 $s_{i,j-1}$에서 거꾸로 진행하여 화제분야 F_{theme}의 분야연상어가 연속으로 출현하고 있는 문의 집합이다. 이 집합의 원소는 Frequency($s_{i,k}$, F_{theme}) ≠ 0과 Frequency($s_{i,j-n-1}$, F_{theme}) = 0을 만족하는 문장집합이다. number(C_i)은 문장집합 C_i의 원소 수 n을 표시한다.

위의 식에 의해 산출된 쇠퇴율을 사용하여 문 $s_{i,j}$에 대한 계속도 α_i를 계산한다. 본 방법에서는 문 $s_{i,i-1}$에서 문 $s_{i,i}$로(해석이 새로운 문으로) 진행할 때, 화제분야 F_{theme}의 계속도가 쇠퇴하고, 문 $s_{i,i}$에서 화제분야 F_{theme}의 분야연상어가 출현하면 화제의 계속성이 상승하였다고 생각한다. 계속도 α_i의 계산 방법을 표시한다.

[계속도 α의 계산]

① $\alpha_j = \alpha_{j-1} + \rho \times Decline_j$

(단, $\alpha_j < 0$의 경우는 $\alpha_j = 0$이 된다.)

② $\alpha_j = \alpha_j + Frequency(s_j, F_{theme})$ [*순서 종료*]

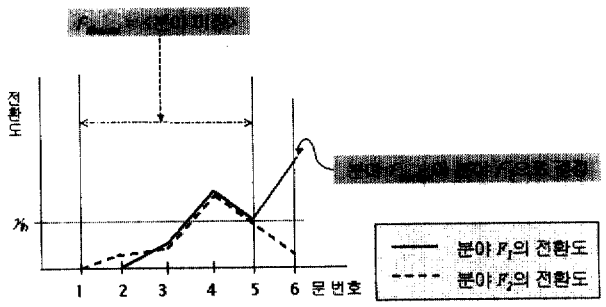
여기서 파라미터 ρ ($0 < \rho < 1$)는 쇠퇴율이 계속도에 영향을 주는 파라미터이다. ρ의 수치가 높으면 쇠퇴율에 영향을 크게 미치게 되어 화제의 계속도가 낮아진다. 그 때문에 화제의 변화가 크거나 변화가 자주 발생하는 문서에 대해 ρ를 높게 설정하면 유용하다. 반대로, ρ를 작게 하면 화제의 연속도가 높아지므로, 화제의 변화가 작은 문서에 유용하다. 문 s_j에서의 분야 F_k에 대하여 전환도 $\beta(F_k)$는 쇠퇴도의 계산식, 또는 위 순서의 F_{theme}을 F_k로 변경하여 계산한다.

4.4 단락의 결정

본 논문의 방법은 각 문장마다 단락분할을 판단한다. 단락결정에 대한 처리를 화제의 출현 판정처리, 전환처리, 계속처리로 나누어 설명한다. 화제분야 F_{theme}, 계속도 α, 그리고 전환도 $\beta(F_k)$의 수치에 의하여 각 처리로 분기한다. 먼저, 화제분야 F_{theme}이 한 가지로 정해지지 않으면 <분야미정>으로 정의한다. 각 문에 대하여 $\beta(F_k)$를 산출해 화제의 출현 판정처리를 수행한다. F_{theme}이 특정한 분야로 결정되어 있으면, α와 $\beta(F_k)$를 계산하여, $\alpha < \beta(F_k)$이면 화제 전환처리를 한다. 반대로 $\alpha > \beta(F_k)$이면 화제 계속처리를 행한다. 아래에 개별적인 처리에 대하여 상세히 설명한다.

4.4.1 화제출현

본 처리에서는 각 분야의 전환도 $\beta(F_k)$로부터 어느 분야가 화제분야 F_{theme}이 되기 쉬운가를 판정한다. 먼저 $\beta(F_k)$가 γ_{th}(임계값)를 넘지 않거나 또는 $\beta(F_k)$가 최대가 되는 분야 F_k가 두 분야 이상 존재하는 경우, F_{theme} = <분야미정>으로 한다. 해석 문을 단락의 후보로 선정하여 스택에 저장한다. 반대로, $\beta(F_k)$가 특정 임계값을 초과하고, 최대가 되는 F_k가 한 분야로 모아지는 경우, F_k를 F_{theme}이라 하고, 단락형성의 후보에서 Frequency($s_{i,j}$, F_{theme}) = 0인 문장 s_j를 선택하여 제외한다. (그림 3)에 표시한 예와 같이 문 s_6에서 $F_{theme} = F_1$이 되고, 문 $s_{i,1}$와 $s_{i,2}$를 선택하여 제외한 $s_{i,3} \sim s_{i,6}$을 분야 F_{theme}의 단락구성 후보문장으로 형성된다.

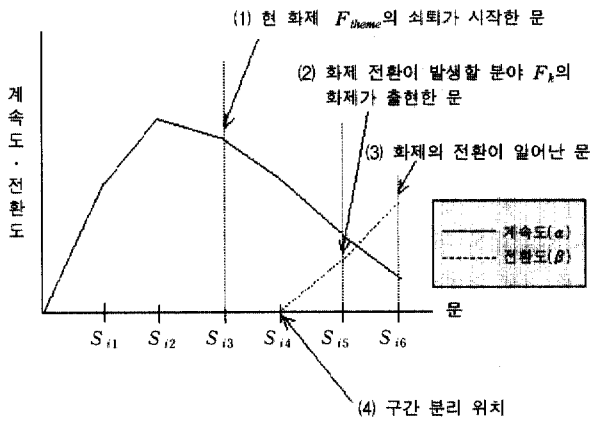


(그림 3) 분야미정으로 스택에 저장된 문장에서 화제분야의 결정

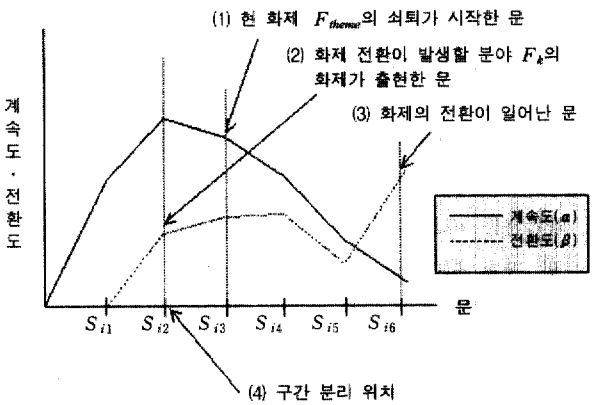
4.4.2 화제전환

화제의 전환이 일어난 경우 인접하는 문장사이에서 구간을 분리할 필요가 있다. 예를 들면, 전환이 일어난 문 $s_{i,j}$에서 새로운 화제가 나타나면, $s_{i,j-1}$을 구간분리위치로 하는 것이 바람직하다. 그러나 화제가 전환되기 이전의 문 $s_{i,j}$까지는 화제가 계속될 가능성이 있기 때문에 스택내의 문장집합에서 처리를 거꾸로 진행하여 구간 분리위치 $s_{i,j}'$을 설정한다.

먼저, 문 $s_{i,j}$에서 한 문장씩 거꾸로 삭제해 가면서 전환되는 분야 F_k의 전환도 $\beta(F_k)$가 최초로 0이 되는 문을 s_j'으로 설정한다. 단, 화제분야 F_{theme}의 계속도가 증가하고 있는 동안은 F_{theme}의 화제분야가 계속하고 있다고 판단하여 $s_{i,j}'$의 계속도의 쇠퇴는 최후에 쇠퇴하기 시작한 다음의 문장이 된다. 따라서 $s_{i,j}'$까지를 전환되는 화제분야의 단락으로 결정하고, 문 $s_{j'+1}$에서 s_j까지를 전환되는 화제분야의 단락 후보로 한다. 예를 들면, (그림 4)(a)의 경우, 문 $s_{i,6}$에서 전환이 일어나고 있으나, 전환도 $\beta(F_k)$가 0 이 되는 문 s_4를 "구간분리위치"로 한다. (그림 4)(b)의 경우, 전환도 $\beta(F_k)$가 0 이 되는 문은 $s_{i,1}$이지만, 문 $s_{i,2}$까지는 계속도가 증가하고 있으므로 구간 분리 위치는 $s_{i,2}$가 된다. 덧붙여, 문 $s_{i,2} \sim s_{i,4}$와 같이 계속도가 감소하기 시작한 부분과 전환도가 증가하기 시작한 부분 중 어느 하나를 우선하여 한 단락으로 결정하는가는 보다 깊은 논의가 필요하지만 이번 연구에서는 전환도의 증가를 우선하기로 하였다.



(a) 화제분야가 쇠퇴한 후에 화제 전환을 야기하는 다른 화제가 출현하는 경우



(b) 새로 전환이 이루어질 화제가 출현 한 후에 화제 분야가 쇠퇴하는 경우

(그림 4) 구간 분리 위치 판정 예

4.4.3 화제계속

화제 계속처리는 계속도 α가 특정 임계값 γ_{th} 이상의 경우에는 화제가 계속되고 있다고 판단하여, 문 $s_{i,j}$를 단락구성의 문장 후보에 추가한다. 만약 α가 α_{th}보다 낮을 경우는 화제가 종료했다고 판단하여 단락 후보에서 Frequency $(s_{i,j}, F_{theme}) = 0$의 문 $s_{i,j}$를 제외한 나머지 문을 Passage (F_{theme})에 추가한다.

5. 실험 및 평가

실험을 위해 <표 3>과 같이 각 분야별·수준별로 구축된 일본어 3,248개의 분야연상어를 구축하였다. 여기서 분야 <스포츠>에 대한 분야연상어가 다른 분야에 비해 훨씬 많은 이유는 타 분야에 비해 <스포츠> 분야의 문서는 인터넷에서 전자화 된 문서를 구해 작업하기 쉽고, 문서 내에 분야를 정확히 결정할 수 있는 분야연상어가 다른 분야에 비해 많이 포함되어 있기 때문이다.

〈표 3〉 일본어 Training Set에서 수집한 각 분야별 분야연상어 수

분야별	분야연상어 수준 1	2	3	4	총합
1. 문화&예술	410	58	28	89	585
	70.0%	9.9%	4.8%	15.3%	100%
2. 비즈니스&경제	144	69	24	39	276
	52.1%	25.0%	8.7%	14.2%	
3. 교육	43	8	1	79	131
	32.8%	6.1%	0.7%	60.4%	
4. 환경문제	100	32	13	66	211
	47.4%	15.2%	6.2%	31.2%	
5. 취미&오락	180	9	0	42	231
	77.9%	3.9%	0%	18.2%	
6. 산업	179	21	0	117	317
	56.5%	6.6%	0%	36.9%	
7. 국제관계	118	10	9	65	202
	58.4%	5.0%	4.5%	32.1%	
8. 자연	126	1	0	43	170
	74.1%	0.6%	0%	25.3%	
9. 스포츠	550	171	135	22	878
	62.6%	19.5%	15.4%	2.5%	
10. 학문	109	10	12	116	247
	44.1%	4.0%	4.9%	47.0%	
수준별 총합	1,959	389	222	678	3,248
	60.4%	11.9%	6.8%	20.9%	100%
평균	195.9	38.9	22.2	67.8	324.8

실험에 사용한 데이터는 주로 아사히신문 CD-ROM 1995~1997년 데이터 컬렉션과 인터넷에서 인간이 수집하여 미리 정해 놓은 분야트리에 의해 분류하였다. 이 문서 데이터에서 무작위로 다섯 분야를 선정하고, 이들 선정된 분야의 문서파일에서 랜덤으로 N행을 선택하여 평가용 문서를 작성하였다. 이 다섯 분야가 혼합된 파일에서 N의 값을 5, 10, 15, 20으로 변화시켜 가며 각각 30 파일을 작성한다. 또한, 분야연상어를 구축할 때 사용한 문서에서 작성한 데이터 셀(Training Set 이라 함)과 분야연상어의 구축에 쓰이지 않은 문서에서 작성한 데이터 셀(Test Set) 등 두 종류로 나누어 실험데이터를 준비하였다.

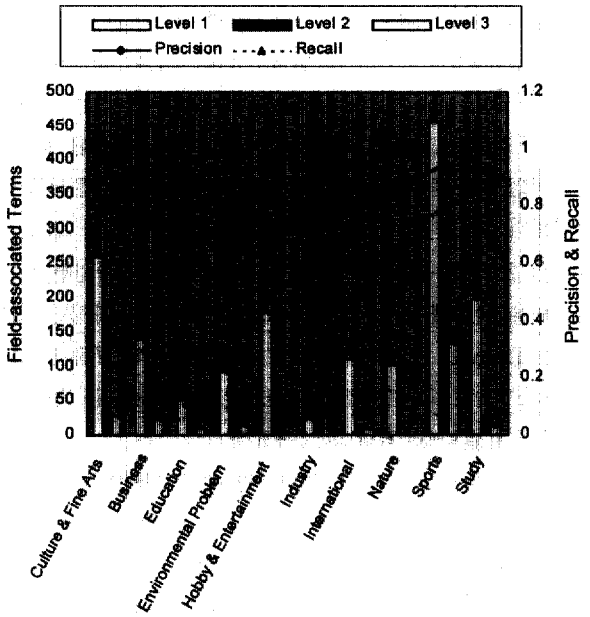
실제로 컴퓨터는 많은 관련되지 않은 단락을 출력한다. 본 시스템이 출력한 단락과 인간이 직접 자신의 상식지식으로 결정한 단락이 어느 정도 일치하는가를 정확율과 재현율을 통해 비교한다. 정확율(P)과 재현율(R)을 아래의 식을 이용하여 계산하였다.

$$P = \frac{P_{correct}}{P_{output}}, \quad R = \frac{P_{correct}}{P_{answer}}$$

여기서, $P_{correct}$은 출력된 단락과 정답 단락이 일치하는 문자수를, P_{output}은 출력 단락의 문자수, P_{answer}는 정답 단락의 문자수를 나타낸다.

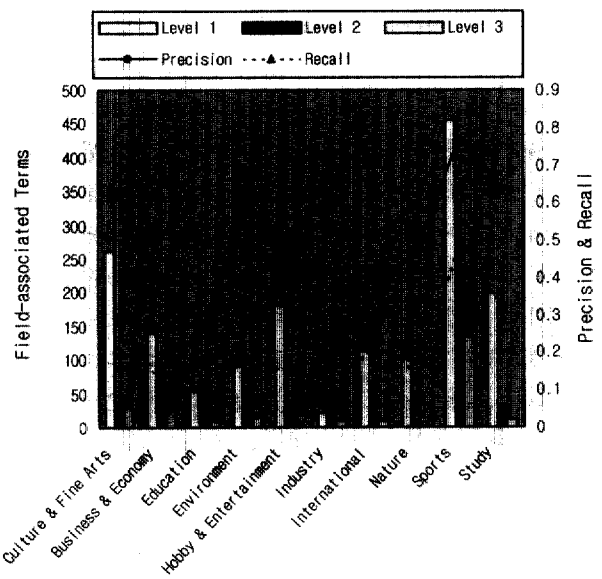
Training Set에 대한 정확도를 (그림 5)[4]에 표시한다. <문화-예술(Culture & Fine Arts)>, <환경문제(Environmental Problems)>, <스포츠(Sports)>, <학문(Study)> 등의 분야는 정확율과 재현율이 전체분야의 평균과 비교하여 높은 이유는 분야연상어의 수가 200을 넘고 있기 때문에 화제분야의 출현, 전환, 계속이 효과적으로 이루어졌다고 생각된다. 반면에, <교육(Education)>과 <산업(Industry)>의 분야에서 재현율이 낮은 이유는 분야연상어의 수가 다른 분야에 비해 상대적으로 적어 화제분야에 해당하는 단락의 출력이 잘 되지 않은 점이 원인이라고 생각된다. 분야연상어 수가 200을 넘고 있는 <비즈니스-경제(Business & Economy)>와 <취미-오락(Hobby & Entertainment)>의 정확율과 재현율이 저하된 이유는 수준 2와 3의 분야연상어의 비율이 높기 때문에 화제분야가 분야마정이 되기 쉽고, 이는 정확율과 재현율 저하의 원인을 초래하였다고 생각된다. 그러나, 각 분야의 평균 정확율은 약 0.82, 재현율은 0.63이 되어 충분히 실용적이며, 본 방법의 유효성이 입증되었다.

4) 다음의 분야 <국제관계>, <자연>, <교육> 등은 구축된 분야연상어의 수가 매우 적기 때문에 실험에서 제외하였으며, 특히, <산업> 분야는 분야연상어 수가 200은 넘지만, 사람에 따라 수준을 크게 다르게 결정하여 이번 실험에서 제외하였다.



(그림 5) 분야연상어 수와 Training Set의 정확율과 재현율

또한, (그림 6)에 Test Set에 대한 실험결과를 표시하였다. Training Set의 정밀도와 비교해 전체적으로 재현율이 저하되어 있다. 이것은 검출된 분야연상어의 수가 극히 감소하였기 때문이다. 그러나 <스포츠>와 같이 질적으로·양적으로 잘 정돈된 분야연상어가 구축되어 있는 분야에 관해서는 정밀도의 저하가 크게 보이지 않아 충분히 실용화 할 수 있다. 따라서, 분야연상어가 구축되기 쉬운 분야나 수준 1의 분야연상어가 많이 존재하는 분야에 대해서는 본 방법은 상당히 높은 정밀도의 단락을 추출할 수 있다.



(그림 6) 분야연상어 수와 Test Set의 정확율과 재현율

6. 결 론

단락검색은 사용자의 검색 질의어에 대해 정확하고 빠르게 동작하고, 검색과 무관한 정보를 빠르게 차단한다. 또한 단락검색은 사용자가 원하는 정보의 존재여부를 빠르게 지시한다. 본 논문에서 제시하는 방법론은 문서내 화제의 계속성과 전환성에 기반한 검색방법이므로 가공되지 않은 자연어문장 형태의 정보를 추출하는 유용한 방법이다. 일반적인 문서의 분야를 분류체계로 정의하고, 이를 트리구조로 표현한 분야트리를 정의하였다. 문서에서 관련된 텍스트를 추출하기 위해 먼저 분야연상어의 연상범위를 나타내는 분야연상어의 수준을 새롭게 제안하였다. 단일 분야연상어의 길이는 의미를 가진 최소한의 단어이고, 개수도 유한하기 때문에 인간이 수집하였다.

결론적으로 본 논문의 방법은 텍스트의 특정 화제분야를 대표하는 실마리로서 분야연상어를 이용하였기 때문에 인간의 두뇌 혹은 인지작용과 유사하게 컴퓨터가 텍스트를 읽어감에 따라 텍스트가 어느 분야에 속하는지 빠르게 판단한다. 또한 단락검색시 화제의 전환성과 계속성을 고려하였기 때문에 동일 분야의 텍스트가 분리되는 현상을 방지하고, 복수분야에 속하는 텍스트의 중복을 제거하는 새로운 단락검색법이다.

향후의 연구과제는 분야연상어간에 공기정보나 격정보를 부여하여 수준 1 이외의 분야연상어가 복수 출현한 경우에도 한가지 뜻으로 문서의 분야를 연상시키는 방법을 고려중이다. 또한, 분야연상어의 상·하위 관계를 정의하고, 문서에서 출현하는 실마리어를 이용하면 단락을 Bottom-up 식으로 요약하는 응용기술도 가능할 것으로 기대한다. 본 연구에서는 각 수준에 대하여 분야연상어의 점수나 파라미터 φ을 한가지 뜻으로 결정하였으나, 코퍼스에서 이 수치를 학습하는 기능도 연구하고자 한다.

참 고 문 헌

[1] Aho, A. V., & Corasick, M. J., "Efficient String Matching : An Aid to Bibliographic Search," Communications of the ACM, Vol.18, No.6, pp.333-340, 1975.

[2] Allan, J., "Relevance Feedback with Too Much Data," Paper Presented at the Proceedings of the 18th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval (SIGIR '95), 1995.

[3] Callan, J. P., "Passage-Level Evidence in Document Retrieval," Paper Presented at the Proceeding of 17th Annual

International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval Research (SIGIR '94), 1994.

[4] Cormack, G. V., Clarke, C. L. A., Palmer, C. R., & To, S. S. L., Passage-Based Refinement (MultiText Experiements for TREC-6), Paper Presented at the Sixth Text REtrieval Conference (TREC-6), 1997.

[5] Daniels, J. J., & Rissland, E. L., "Locating Passages Using a Case-Base of Excerpts," Conference on Information and Knowledge Management, Paper Presented at the Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management, 1998.

[6] Dozawa, T., (Editor) "Innovative Multi-Information Dictionary, Imidas '99," Annual Series, Zueisha Publication Co., 1999, (in Japanese).

[7] Fuketa, M., Lee, S., Tsuji, T., Okada, M., & Aoe, J., "A Document Classification Method by Using Field Association Words," An International Journal of Information Sciences, Elsevier Science, Vol.126, No.1-4, pp.57-70, 2000.

[8] Hearst, M. A., & Plaunt, C., "Subtopic Structuring for Full-Length Document Access," Paper Presented at the Proceedings of 16th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval Research (SIGIR '93), 1993.

[9] Hess, M., "Deduction over Mixed-Level Logic Representations for Text Passage Retrieval," Paper Presented at the Proceedings of the 1996 International Conference on Tools with Artificial Intelligence (TAI '96), 1996.

[10] Hoenkamp, E., & Groot, R., "Finding Relevant Passages Using Noun-Noun Compounds : Coherence vs. Proximity," Paper Presented at the Twenty-Third Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval (SIGIR 2000), 2000.

[11] Iwayama, M., & Tokunaga, T., "Probabilistic Passage Categorization and its Application," Journal of Natural Language Processing, Vol.6, No.3, pp.181-198, 1999, (in Japanese).

[12] Kaszkiel, M., & Zobel, J., "Passage Retrieval Revisited," Paper Presented at the Proceeding of 20th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval Research (SIGIR '97), 1997.

[13] Kaszkiel, M., Zobel, J., & Sacks-Davis, R., "Efficient Passage Ranking for Document Databases," ACM Transactions on Information Systems, Vol.17, No.4, pp.406-439, 1999.

[14] Knaus, D., Mittendorf, E., & Schauble, P., "Improving a Basic Retrieval Method by Links and Passage Level Evidence," Paper Presented at the Third Text REtrieval Conference (TREC-3), 1994.

[15] Knaus, D., Mittendorf, E., Schauble, P., & Sheridan, P., "Highlighting Relevant Passages for Users of the Interactive SPIDER Retrieval System," Paper Presented at the Fourth Text REtrieval Conference (TREC-4), 1995.

[16] Kurohashi, S., Shiraki, N., & Nagao, M., "A Method for Detecting Important of a Word Based on Its Density Distribution in Text," Paper Presented at the Transactions of Information Processing Society of Japan, Vol.38, No.4, pp.845-854, 1997, (in Japanese).

[17] Lee, S., Koyama, M., Mizobuchi, S., Uchibayashi, K., Kawano, F., Komatsu, T., & Aoe, J., "Cross-Language Multi-Media Information Retrieval System : BOSS," Paper Presented at the 18th International Conference on Computer Processing of Oriental Languages (ICCPOL '99), 1999.

[18] Melucci, M., "Passage Retrieval : A Probabilistic Technique, An International Journal of Information Processing and Management," Vol.34, No.1, pp.43-63, 1998.

[19] Mittendorf, E., & Schauble, P., "Document and Passage Retrieval based on Hidden Markov Models," Paper Presented at the Proceeding of 17th Annual International ACM Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval Research (SIGIR 94), 1994.

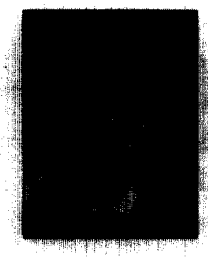
[20] Mizuno, H., Kise, K., & Matsumoto, K., "Linking Figures and Tables to Their Expository Texts Using Word Density Distributions and Their Biases," Paper Presented at the Transactions of Information Processing Society of Japan, Vol.40, No.12, pp.4400-4403, 1999, (in Japanese).

[21] Mochizuki, H., Makoto, I., & Okumura, M., "Passage-Level Document Retrieval Using Lexical Chains. Journal of Natural Language Processing," Vol.6, No.3, pp.101-126, 1999, (in Japanese).

[22] O'Connor, J., "Retrieval of Answer-Sentences and Answer-Figures from Papers by Text Searching," An International Journal of Information Processing & Management, Vol.11, No.5/7, pp.155-164, 1975.

[23] Tsuji, T., Nigazawa, H., Okada, M., & Aoe, J., "Early Field Recognition by Using Field Association Words," Paper Presented at the Proceedings of the 18th International Conference on Computer Processing of Oriental Language (ICCPOL '99), 1999.

[24] Yasutake, M., Koyama, Y., Yoshimura, K., & Shudo, K., "Kana-to-Kanji Conversion Systems Based on Large Scale Collocation Data," Paper Presented at the Proceedings of the 1999, 18th International Conference on Computer Processing of Oriental Language (ICCPOL '99), 1999.



이 상 곤

e-mail : samuel@jeonju.ac.kr

1996년 전북대학교 컴퓨터과학과(학사)

1998년 전북대학교 전산통계학과(이학석사)

2001년 일본 도쿠시마대학교 지능정보공학과(공학박사)

2001년~2002년 원광대학교 음성정보 기술산업 지원센터 연구원

2002년~현재 전주대학교 정보기술컴퓨터공학부 컴퓨터공학전공 전임강사

관심분야 : 한국어 정보처리, 한글공학, 정보검색, 문서분류