이용자 질의 기반 단락추출에 관한 연구

A Study on Extracting Passage by Users' Query

최상희, 연세대학교 대학원 문헌정보학과, shchoi@lis.yonsei.ac.kr

Sang-Hee Choi, Dept. of Lib. & Info. Sci., Graduate School of Yonsei University

단락은 문서의 세부 주제를 담고 있는 중요한 단위이다. 이 연구에서는 이용자 다양한 질의에 맞추어 동적으로 주제단락을 추출하여 이용자가 찾고자 하는 정보를 제공하는 방안을 고찰하였다. 추출된 단락의 질의응답성능을 분석, 평가한 결과 복수문서 환경에서 순차적 단락확장 기법으로 추출된 단락이 이용자 질의에 가장 적합한 정보를 추출하는 것으로 나타났다.

1. 서론

이용자의 질의는 매우 다양하며 예측하기 어려운 특성이 있다. 따라서 이용자 질의에 따라 단락을 추출한다는 것은 상당히 어려운 과제이다. 그러나 이용자 질의에 맞추어 질의에 가장 적합한 정보를 그때그때 추출할 수 있다면 문서전체를 보기보다는 질의에 맞는 일부 정보만을 보고 싶어 하는 이용자에게는 매우 효율적인 정보제공 방안이 될 수 있다. 이 연구에서는 단일문서와 복수문서 환경에서 이용자 질의에 따라 문서에서 단락을 동적으로 추출한 후, 단락추출 기법별로 이용자 질의에 대하여 응답하는 성능을 평가하였다.

2. 단락의 유형과 추출 과정

2.1 단락의 생성시기에 따른 유형

단락(passage)은 대부분 자동으로 분할될 수 있는 일정량의 연결된 정보로서 문헌으로부터 추출한 연속된 텍스트이다(Kaszkiel, and Zobel 2001).

단락이 생성되는 시기에 따른 유형을 살펴보자면 문서가 처음 입력될 때 주제나 특정요소에 따라 규정해 놓는 정적단락이 있고 이용자 질의나 외부조건에 따라 변화하는 동적단락이 있다(Callan 1994, 김정하 2001).

이용자 질의를 중심으로 단락을 분석해보고자 하면 단락의 생성시기는 매우 중요한 요소 중 하나이다. 이용자 질의는 문서가 생성된 이후에 매우 다양하게 나타나기 때문에 이에 맞추어 효율적이고 경제적으로 단락을 생성하는 상당히 어려운 과제이다.

정적단락은 주로 문서 내 주제에 따라 미리 분할되어 있는 경우에 해당하므로 질의가 미리 준비해 놓은 주제단락에 대응하지 않을 경우 정적단락은 질의에 딱 맞는 정보를 제공하지 못 할 수 있다. 정적단락의 특성은 한번 문서의 주제를 분석하여 단락을 분할해 놓으면, 생성된 단락은 변화하지 않고 따라서 단락의 주제도 고정된다는 것에 있다. 반면 동적단락은 단락이 이용자 질의에 따라 그때그때 생성된다. 즉 단락주제를 이용자 질의에 맞추어 질의에 가장 적합한 정보를 문서에서 검색한 후 그 검색 결과를 중심으로 정보를 확장하여 단락을 추출하는 것이다. 동적단락은 이용자 질의처럼 동적으로 변화하는 특성을 가진 요소에 적합하다. 정적단락이 정보의 분류 및 분할에 기반을 둔 접근방식이라면 동적단락은

텍스트 분할의 개념보다는 정보검색과 확장, 추출에 기반을 둔 접근방식이다.

2.2. 단락확장을 기반으로 한 동적단락추출

동적단락 추출을 위해서 사용되는 기법 중 하나인 문장기준의 단락확장은 한 문장을 중심으로 하여 주제적으로 연결되어 있는 주변 문장을 분석하여 연관된 부분으로 확장하여 추출하는 것이다. 즉, 질의와 가장 맞는 문장을 검색한 후 이 문장을 중심으로 이웃문장과의 관계를 분석하여 연관성이 높은 이웃문장을 단락으로 확보하는 것이다.

동적으로 단락을 확장해나가는 방식으로는 개념확장 활성화 기법이 있다. 개념확장 활성화 기법은 인공지능기반 시스템에서 초기노드와 연결된 노드를 따라 항해하면서 적합한 정보를 검색해나가는 기법이다. 개념확장 활성화 기법 중 비교적 좋은 성능을 나타내고 있는 bnb (branch-and-bound search)기법은 개념확장이 진행되는 동안 최단 경로를 찾기 위한 방법이다(노영희 2000).

bnb 확장 활성화 기법은 확장 중심노드에서 가장 유사한 방향으로 노드를 추적해가는 방식이므로, 특정 문장을 중심으로 가장 유사도가 높은 문장으로 확장해나갈 수 있는 방안을 제공할 수 있다. 예를 들면 이용자 질의의 주제를 포함한 문장을 기준으로 bnb 확장 활성화 방식을 적용하면 주변 문장으로 확장시키는 과정을 통해 문서 집단 내에서 질의에 답이 될 만한 주요 단락을 추출할 수 있는 것이다.

3. 동적단락 추출을 위한 단락확장 기법

3.1 순차적 bnb 단락확장 기법

순차적 bnb 단락확장 기법은 단락확장 기준이 되는 중심문장들을 가지고 주변문장과 비교했을 때, 비교된 모든 문장에서 최단거리에 있는 주변문장을 하나만 선정하는 방식이다. 즉, 주변단락으로 확장할 때 중심문장에서 주변 문장까지의 유사도를 모두 기억시킨 후 그중 유사도가 가장 높은 1개만을 선택하여 확장하는 방식이다. 이때 확장결과로 포함된 단락은 2차확장 단계의 단락확장 중심으로 포함된다. 2차 확장에서는 새 기준에 따라 다시 모든 경로를 비교하고 2차 단락확장 결과로 1개를 선택하는 순환과정을 통해서 단락확장이 이루어진다. 그림 1은 3개의 문장을 중심으로 순차적 bnb 단락확장을 하는 예이다.

3.2 병렬적 bnb 단락확장 기법

병렬적 bnb 단락확장은 순차적 단락확장이 주심문장에서 가장 최단 경로만을 포함시켜 확장하는 방식을 택하는 것에 반해 중심문장 각각에 대하여 2차 확장이 이루어지는 것이다. 병렬적 bnb 단락확장 과정은 그림 2와 같다. 병렬적 bnb 단락확장에서는 중심문장 3개를 중심으로 앞뒤 단락들과 유사도를 산출하여 각 중심 문장별로 유사도가 큰 순으로 순위를 매겨 중심문장 당 1개씩 유사도가 높은 단락을 추출한다.

그림에서 나타났듯이 순차적 단락확장에서는 1차 확장시 3개의 문장과 가장 최단경로인 문장만을 선택하여 확장하므로 1차 확장 결과 4개의 문장이 선정되는 것이고 병렬적 bnb 단락확장에서는 3개의 문장을 기점으로 각각의 최단경로를 다 확장대상문장으로 포함시키므로 1차 확장 결과 6개의 문장이 되는 것이다.

4. 단락확장기법별 응답성능 평가실험

4.1 실험개요

동적단락 추출에 적용된 기법은 bnb 확장 활성화 방식을 기반으로 한 순차적 bnb 단락확장 기법과 병렬적 bnb 단락확장 기법이다. 두 기법은 단락을 추출하는 범위를 단일문서 환경과 복수문서 환경, 두 가지로 차별을 두어 문서환경에 따른 단락확장 성능도 비교되었다.

단락확장 중심문장과 유사도 총순위 ↓

[1차 확장]

<첫째 중심문장 앞뒤 단락>

둘째 '경제적 기준'은 신규 회원국이 시장 경제를 작동시켜야 하고, ~~~~ 회원국으로서 의무를 이행할 수 있는 능력을 가져야 한다. — 5위

첫째 중심문장: 1998년 3월 EU는 헝가리, 폴란드, 체코, 슬로베니아, 에스토니아, 사이프러스 등 6개국을 우선 가입 대상국으로 선정하고 가입 협상을 개시했다.

1999년 12월 헬싱키 EU 정상회담에서 불가리아, 루마니아, 슬로바키아,~~~~2월부터 가입 협상이 시작됐다. 개혁 실적, EU 15개국 요구에 못미치고 있다. — 3위

<둘째 중심문장 앞뒤 단락>

이처럼 미국이 EU의 최대 시장이기 때문에 철강과 농산물 등에서 미국과 EU의 ~~~~또 다시 미국과 EU의 관계가 악화일로를 달리고 있다. — 4위

둘째 중심문장: EU는 새해 상반기 안에 폴란드, 헝가리, 체코, 슬로베니아, 슬로바키아, 발트 3국, 몰타, 사이프러스 등 10개국을 회원으로 받아들인다.

따라서 25개 회원국으로 늘며 인구수가 4억5000만명~~~·세계 수출의 19%를 차지하고 있다(미국 16%, 일본 9%). — 6위

<셋째 중심문장 앞뒤 단락>

두 번째는 특히 젊은이들에게 큰 고통을 주고 있는 높은 실업률~~~~ 분야는 투자환경의 개선이다. — 1위

셋째 중심문장: 현재 체코, 헝가리 등 주변국들은 외국인 투자를 위하여 정부에서 적극 나서고 있는 반면 폴란드는 아직까지 비효율적인 관료체제로 인하여 투자유치에 소홀한 면이 많다.

폴란드가 이러한 어려움을 ~~~~지도가 바뀐다. — 2위

1차 확장 결과 ↓ 중심문장과 유사도 비교결과 제일 유사도가 높은 단락 1개만 선택

기본 중심문장:

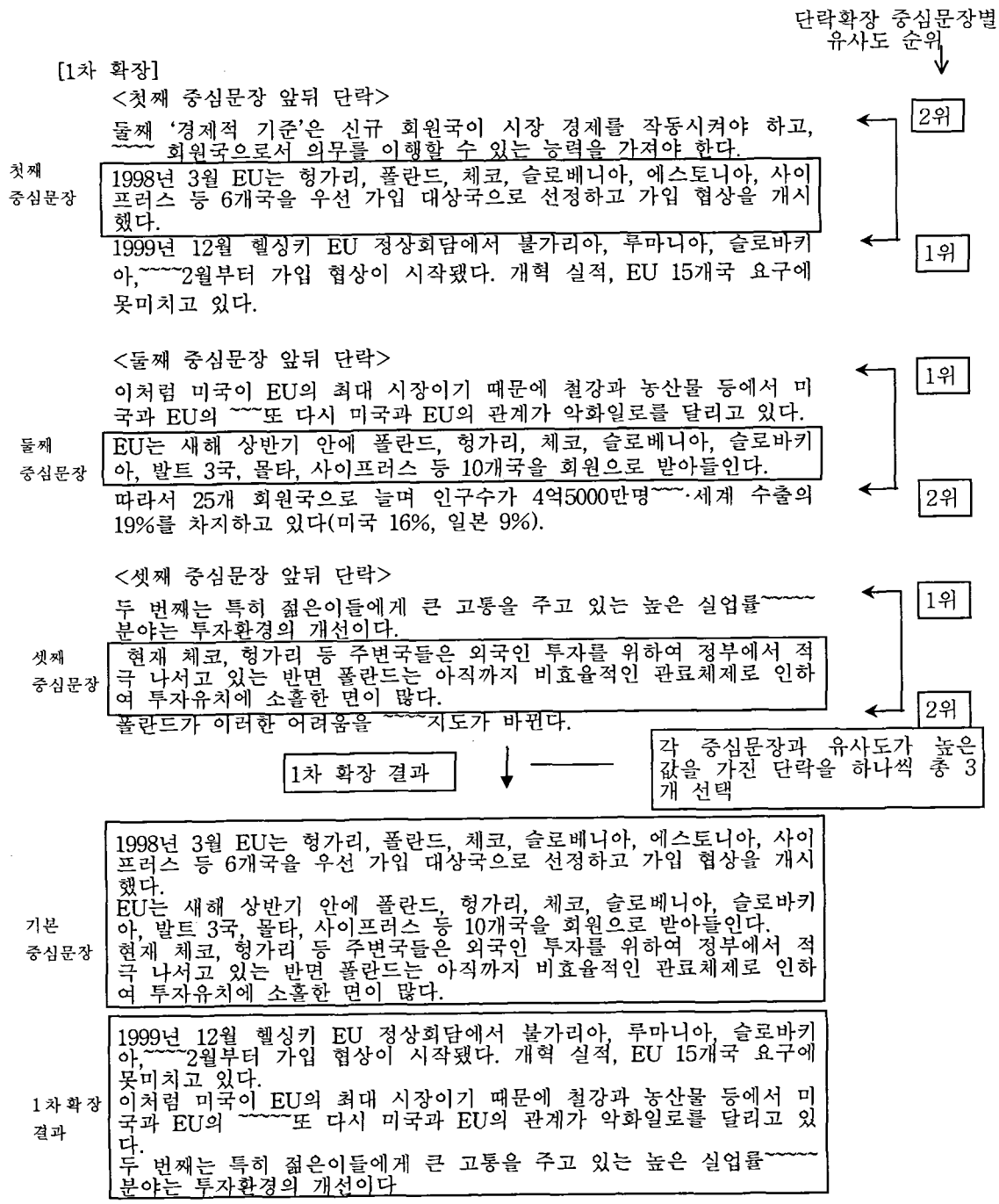
1998년 3월 EU는 헝가리, 폴란드, 체코, 슬로베니아, 에스토니아, 사이프러스 등 6개국을 우선 가입 대상국으로 선정하고 가입 협상을 개시했다.

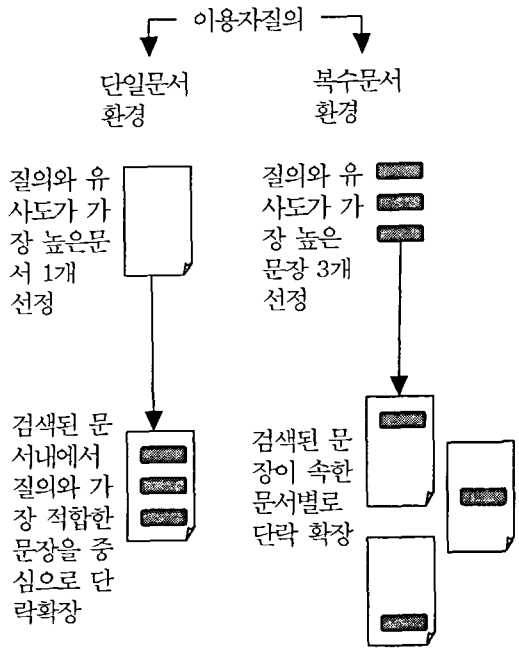
EU는 새해 상반기 안에 폴란드, 헝가리, 체코, 슬로베니아, 슬로바키아, 발트 3국, 몰타, 사이프러스 등 10개국을 회원으로 받아들인다.

현재 체코, 헝가리 등 주변국들은 외국인 투자를 위하여 정부에서 적극 나서고 있는 반면 폴란드는 아직까지 비효율적인 관료체제로 인하여 투자유치에 소홀한 면이 많다.

1차 확장 결과: 두 번째는 특히 젊은이들에게 큰 고통을 주고 있는 높은 실업률~~~~ 분야는 투자환경의 개선이다

<그림 1> 순차적 단락확장의 과정-사례


단락확장 중심문장별
유사도 순위
[1차 확장]
<첫째 중심문장 앞뒤 단락>
둘째 '경제적 기준'은 신규 회원국이 시장 경제를 작동시켜야 하고, ~~~~ 회원국으로서 의무를 이행할 수 있는 능력을 가져야 한다.
2위
첫째 중심문장
1998년 3월 EU는 헝가리, 폴란드, 체코, 슬로베니아, 에스토니아, 사이프러스 등 6개국을 우선 가입 대상국으로 선정하고 가입 협상을 개시했다.
1999년 12월 헬싱키 EU 정상회담에서 불가리아, 루마니아, 슬로바키아,~~~~2월부터 가입 협상이 시작됐다. 개혁 실적, EU 15개국 요구에 못미치고 있다.
1위
<둘째 중심문장 앞뒤 단락>
이처럼 미국이 EU의 최대 시장이기 때문에 철강과 농산물 등에서 미국과 EU의 ~~~또 다시 미국과 EU의 관계가 악화일로를 달리고 있다.
1위
둘째 중심문장
EU는 새해 상반기 안에 폴란드, 헝가리, 체코, 슬로베니아, 슬로바키아, 발트 3국, 몰타, 사이프러스 등 10개국을 회원으로 받아들인다.
따라서 25개 회원국으로 늘며 인구수가 4억5000만명~~~·세계 수출의 19%를 차지하고 있다(미국 16%, 일본 9%).
2위
<셋째 중심문장 앞뒤 단락>
두 번째는 특히 젊은이들에게 큰 고통을 주고 있는 높은 실업률~~~~~ 분야는 투자환경의 개선이다.
1위
셋째 중심문장
현재 체코, 헝가리 등 주변국들은 외국인 투자를 위하여 정부에서 적극 나서고 있는 반면 폴란드는 아직까지 비효율적인 관료체제로 인하여 투자유치에 소홀한 면이 많다.
폴란드가 이러한 어려움을 ~~~~지도가 바뀐다.
2위
각 중심문장과 유사도가 높은 값을 가진 단락을 하나씩 총 3개 선택
1차 확장 결과
기본 중심문장
1998년 3월 EU는 헝가리, 폴란드, 체코, 슬로베니아, 에스토니아, 사이프러스 등 6개국을 우선 가입 대상국으로 선정하고 가입 협상을 개시했다.
EU는 새해 상반기 안에 폴란드, 헝가리, 체코, 슬로베니아, 슬로바키아, 발트 3국, 몰타, 사이프러스 등 10개국을 회원으로 받아들인다.
현재 체코, 헝가리 등 주변국들은 외국인 투자를 위하여 정부에서 적극 나서고 있는 반면 폴란드는 아직까지 비효율적인 관료체제로 인하여 투자유치에 소홀한 면이 많다.
1차 확장 결과
1999년 12월 헬싱키 EU 정상회담에서 불가리아, 루마니아, 슬로바키아,~~~~2월부터 가입 협상이 시작됐다. 개혁 실적, EU 15개국 요구에 못미치고 있다.
이처럼 미국이 EU의 최대 시장이기 때문에 철강과 농산물 등에서 미국과 EU의 ~~~~또 다시 미국과 EU의 관계가 악화일로를 달리고 있다.
두 번째는 특히 젊은이들에게 큰 고통을 주고 있는 높은 실업률~~~~~ 분야는 투자환경의 개선이다



<그림 3> 문서환경별 단락확장 과정

단일문서 환경에서의 단락확장은 그림 3과 같이 먼저 이용자의 질의에 가장 적합한 문서를 검색한 후, 검색된 문서에서 질의에 가장 적합한 문장 3개를 찾아 그 문장들을 중심으로 단락을 확장하는 방식이다.

복수문서 환경에서 단락확장은 이용자의 질의에 적합한 문장을 여러 문서에서 추출한 후 질의에 적합한 문장 3개를 중심으로 단락을 확장하는 방식이다. 따라서 단락확장은 동시에 여러 문서에서 이루어지게 된다.

실험문서 집단은 주간조선과 주간한국 기사 3000건이며 질의확장에 적용된 질의는 총 10개이다. 각 질의의 평균 질의어 수는 5.1이다.

질의와 유사도를 비교하는 용어가중치로는 이진빈도와 역문장 가중치(isf)를 사용하였고 유사계수는 내적계수를 사용하였다. 역문장 가중치는 역문헌 빈도를 응용한 것이다.

$$\text{역문장빈도(isf)} = 1 + bg2\frac{NS}{sf}$$

NS : 문헌그룹내 문장의 총수
sf : 문장빈도

4.2. 실험결과 분석

순차적 bnb와 병렬적 bnb 방식으로 단락확장을 하여 요약문을 생성한 결과, 각 문서환경에서 순차적 bnb 기법이 병렬적 bnb 기법보다 질의에 적합한 문장을 효과적으로 추출하는 것으로 나타났다.

평가에 적용된 척도는 단락내 문장 정확률과 중복이다.

$$\text{단락 문장 정확률} = \frac{\text{질의에 적합한 문장 수}}{\text{단락내 총 문장 수}}$$

$$\text{단락 문장 중복률} = \frac{\text{정보가 중복되는 문장 수}}{\text{단락내 총 문장 수}}$$

표 1에서 나타났듯이 단일문서 환경에서 순차적 bnb 단락확장 기법은 병렬적으로 단락확장을 한 것 보다 질의에 적합한 문장을 추출하는 성능이 좋은 것으로 나타났다. 또한 추출된 단락간 중복성을 비교하였을 때에도 순차적 bnb 단락확장 기법이 정확하면서도 중복되는 정보는 적은 단락을 추출하는 것으로 밝혀졌다. 복수문서 환경에서도 순차적 bnb 단락확장은 병렬적 bnb 단락확장 기법보다 질의를 기반으로 단락을 추출하는데 효과적인 것으로 나타났다.

<표 1> 문서환경별 단락확장 기법 성능비교

	단락확장 기법	단락 문장 정확률	단락 문장 중복률
단일문서 환경	순차적 bnb	0.621	0.362
	병렬적 bnb	0.584	0.385
복수문서 환경	순차적 bnb	0.675*	0.203*
	병렬적 bnb	0.590	0.270

각 문서환경별 나타난 결과를 비교하여 보면 복수문서 환경에서 순차적 bnb 단락확장의 정확률은 0.675로 병렬적 bnb 단락확장 0.590보다는 14% 개선된 것으로 나타난 것에 반해

단일문서 환경에서는 순차적 bnb 단락확장의 정확률이 0.621로 병렬적 단락확장의 0.584 보다 6% 성능이 좋은 것으로 나타나 실제 큰 차이를 보이고 있지 않았다. 또한 표 1을 보자면 단일문서 환경에서는 단락문장의 중복률에서도 기법간 차이가 복수문서 환경보다는 두드러지게 나타나고 있지 않았다.

특히 단일문서 환경에서는 단락을 확장하는 범위가 질의와 적합한 한 문서로 제한되어 있다는 한계 때문에 단락확장시 중심문장에서 확장되어 추출되는 부분들이 중복된 내용을 다루는 경우가 많았다. 일반적으로 확장되는 단락은 확장 중심문장과 중복되는 정보를 가지고 있을 확률이 높다. 그러나 단일문서 환경에서는 특히 확장되는 단락이 확장 중심문장뿐만 아니라 같이 확장되나 또는 이미 확장되어 있는 단락과의 주제 중복성도 높은 것으로 나타났다. 반면 복수문서 환경에서는 확장된 단락이 이미 추출된 단락과 중복되거나 다른 단락을 확장하는데 사용하는 중심문장과 중복되는 경우도 많지 않았기 때문에 결과적으로 단일문서 환경에 비해 중복성이 적은 것으로 여겨진다.

5. 결론

순차적 bnb 단락확장은 단일문서와 복수문서 두 환경에서 모두 병렬적 bnb 단락확장보다 좋은 평가를 받았다. 특히, 순차적 bnb 단락확장은 복수의 문서들과 같이 단락추출 대상의 범위가 넓어질 때에 질의에 정확한 정보를 찾아 단락으로 추출해내는 성능이 뛰어난 것으로 나타났다. 또한 순차적 bnb 단락확장은 추출된 단락내 중복 정보 비율이라는 측면에서도 병렬적 bnb 단락확장기법보다 효과적인 것으로 나타났다.

병렬적 bnb 단락확장을 기반으로 단락추출은 적합정보를 추출할 때 같은 주제를 집중적으로 추출해낸다는 성향을 나타내었지만 순차적 bnb 단락확장 기법은 유사한 주제로 확장해나가면서도 상대적으로 중복되는 주제보다는 주변 주제나 유사 주제로 확장해나가는 성향이 있는 것으로 분석되었다.

이상의 실험결과를 통해서 이 연구에서는 이용자 질의에 적합한 단락을 추출하는 가장 효율적인 방안으로 순차적 bnb 단락확장 기법을 제안한다.

참고문헌

김정하. 2001. 이용자 중심 요약문 생성에 관한 실험적 연구. 석사학위 논문, 연세대학교, 문헌정보학과.

노영희. 2000. 의미망 지식베이스를 이용한 개념기반 정보검색 기법에 관한 실험적 연구. 박사학위 논문, 연세대학교, 문헌정보학과.

Callan, J. P. 1994. "Passage-level Evidence in Documentation Retrieval". *In Proceedings of the 17th Annual International ACM-SiGIR Conference on Research and Development in Information Retrieval,* 302-309.

Clark , C. L. A., Cormack, G. V., Kisman, D. I. E and Lynam. T. R. 2001. "Question Answering by Passage Selection"
<http://citeseer.ist.psu.edu/454871.html>

Kaszkiel, M., and Zobel, J. 2001. "Effective Ranking with Arbitrary Passages". *Journal of the American Society for Information Science and Technology,* 52(4): 344-364.