복합명사 분리 색인 방법이 문서 클러스터링에 미치는 영향 분석

양명석[0], 최성필
한국과학기술정보연구원
{msyang[0], spchoi}@kisti.re.kr

An Analysis of the Hierarchical Agglomerative Clustering based on various Compound Noun Indexing Method

Myung-Seok Yang, Sung-Pil Choi[0]

요 약

본 논문에서는 복합명사에 대한 색인 방법을 다각적으로 적용하여 계층적 결합 문서 클러스터링 시스템의 결과를 분석하고자 한다. 우선 한글 색인 엔진과 HAC(Hierarchical Agglumerative Clustering) 엔진에 대해서 설명하고 한글 색인 엔진에서 제공되는 세가지 복합명사 분석 모드에 대해서 설명한다. 또한 구현된 클러스터링 엔진의 특징과 속도 향상을 위한 기법 등을 설명한다. 실험에서는 다양한 요소를 가지고 클러스터링된 문서 집합에 대한 분석 결과를 보인다. 실험 결과에 대한 분석에서 복합명사에 대한 색인 방법이 문서 클러스터링의 결과에 직접적인 영향을 준다는 것을 보여준다.

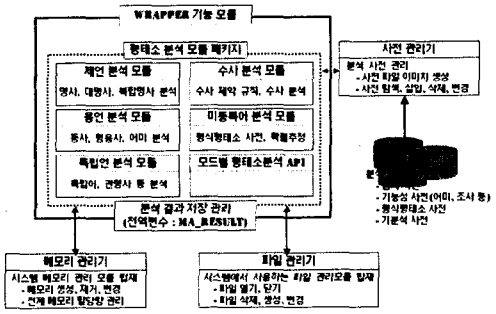
1. 서 론

문서 클러스터링이란 비슷한 성향을 나타내는 문서들간의 관계를 분석하여 하나의 군집으로 모으는 것이다. 문서에 대한 유사성 판단을 위해서는 문서를 색인하여 색인 어휘를 추출하고 추출된 색인어를 바탕으로 문서 벡터를 구성해야 한다. 특수한 경우를 제외하고는 대부분의 시스템들은 명사 및 명사 상당어구를 색인어로 추출하게 된다.

이 논문에서는 색인어로서 명사를 추출할 때, 특히 복합명사에 대한 분석 방법이 전체 클러스터링의 결과에 미치는 영향에 대해서 논하고자 한다. 우선 2장에서는 본 시스템에서 사용된 한글 자동 색인 시스템에 대해서 설명하고, 제공되는 3가지 복합명사 분석 방법에 대해서 설명한다. 3장에서는 이 논문에서 구현된 계층적 결합 클러스터링 엔진에 대해서 설명하고 4장에서는 다양한 실험을 통해서 3가지 복합명사 분석 방법이 결과로 나온문서 클러스터의 특징에 어떠한 영향을 주는지를 분석한다.

2. 한국어 자동 색인 시스템

이 논문에서 사용한 한국어 자동 색인 시스템의 구조는 [그림 1]과 같다. [그림 1]에서 보는 바와 같이 기능 모듈들이 서로 유기적으로 결합되어 있는 형태이므로 시스템의 커스터마이징이 쉽고 다양한 기능들을 새로이 추가할 수 있다. 이 논문에서는 색인어 추출을 위해서 품사 태거를 사용하지 않는다. 따라서 [그림 1]의 형태소 분석 시스템이



[그림 1] 자동 색인 시스템의 전체 구조도

바로 색인 시스템이 된다. 물론 정확한 색인을 위해서는 품사 태거를 사용하는 것이 정설로 되어 있으나, 복합명사의 색인 방법을 다양하게 지정해야 하고 시스템의 속도 등을 고려하여 형태소 분석 결과에서 바로 명사 및 명사 상당어구만을 추출하는 식의 자동 색인을 수행하였다.

일반적인 복합명사에 대한 색인 방법은 복합명사를 무조건 단위명사로 나누는 방법이다. 그러나 복합명사 그 자체가 하나의 의미를 가지면서 단위명사로 분리되면 그 의미를 상실하는 경우가 많다. 그 예로 "대우전자"는 "대우"와 "전자"로 분리되면 "사장대우", "잘 대우하다"와 같은 뜻의 "대우"와 "전자, 후자"와 같은 경우로 사용되는 "전자"로 나뉘진다. 그러나 "대우전자"는 하나의 고유명사로서 다른 의미구조를 가진다. 이렇게 무조건 복합명사를 단위명사로 분리하게 되면 클러스터링을 위한 문서 벡터 공간을 형성할 때, 위와 같은 단어가 포함된 문서를 구별하지 못하므로 문서 고유의 특징을 상실하게 된다. 또한 본 논문에서는 품사 태거를 사용하지 않으므로 한 복합명사에 대한 결과가 여러 가지일 경우에 이들 단위명사를 모두 색인어로 추출하게 된다.

복합명사 분석 방법은 크게 3가지이다.(CAI, CAN, CAOI) 이들 분석 방법에 대한 설명은 아래 표에 나타나 있다.

[표 1] 복합명사 색인 방법

CN=(N_1,N_2,...N_n) : 분석결과 나온 모든 단위명사들 집합		
종류	추출 색인어	설명
CAI	N_1,N_2,...N_n,CN	복합명사 자체를 색인어에 포함
CAN	N_1,N_2,...N_n	단위명사만을 색인어에 포함
CAOI	CN	복합명사 자체만을 색인어에 포함

위 3가지 방법에 의한 색인을 수행하고 각각의 방법에 따라서 색인 결과를 클러스터링 엔진으로 넘겨주게 된다.

3. HAC기반 문서 클러스터링 엔진

본 논문에서 개발된 문서 클러스터링 엔진의 구조는 [그림 2]와 같

[그림 2] 문서 클러스터링 엔진의 구조

다.

문서 클러스터링은 크게 2단계로 이루어진다. 첫 번째 단계에서는 입력 문서를 색인하고 색인어를 추출하여 데이터베이스에 각종 통계정보를 저장하고 문서 벡터를 저장하게 된다. 두 번째 단계에서는 생성된 문서 벡터 정보를 이용하여 문서간 유사도와 클러스터간 유사도를 중심으로 클러스터링을 수행하게 된다. 문서 벡터를 구성하는 각 단어에 대한 가중치 계산은 일반적으로 많이 사용되는 로그 tf*idf를 사용하였다.

$$w_{d,t} = (K + (1-K)\frac{f_{d,t}}{\max_t f_{d,t}}) \times \log\frac{N}{f_t} \quad \text{---------- (1)}$$

여기서 $f_{d,t}$는 문서 d내에서 단어 t의 빈도이고 f_t는 단어 t의 전체 단어 빈도이다. K값은 0으로 지정하였다. 문서간의 유사도는 코사인 계수를 그대로 적용하였다.

$$Sim(D_i, D_j) = \frac{D_i \cdot D_j}{|D_i \| D_j|} = \frac{1}{W_i W_j}\sum_{t=1}^{n} w_{i,t} \cdot w_{j,t}$$

$$W_i = \left(\sum_{t=1}^{n} w^2_{i,t}\right)^{1/2}, W_j = \left(\sum_{t=1}^{n} w^2_{j,t}\right)^{1/2} \quad \text{----- (2)}$$

클러스터간 유사도 측정은 군집 평균 연결(group average link)방법을 이용한다. 클러스터간 유사도가 특정 임계치보다 큰 두 클러스터가 하나로 병합되게 된다. 이런 병합은 클러스터 집합 내의 모든 클러스터쌍이 임계치보다 낮을 때까지 반복되게 된다.

$$SimC(C_i, C_j) = \frac{1}{|C_i \| C_j|}\sum_{d_i \in C_i}\sum_{d_j \in C_j} SimD(d_i, d_j) \quad \text{---- (3)}$$

클러스터링 과정에서 문서 집합 내의 모든 문서에 대해서 유사도를 측정해야하는 부담을 덜기 위해서 문서 벡터 생성시에 같은 색인어를 공유하는 문서를 선택하여 연관 문서 연결 DB에 저장하게 된다. 이렇게 하면 클러스터링 과정에서 특정 문서에 대한 연관 문서만에 대해서만 유사도 측정을 하면 되므로 유사도가 특정 임계치 이하의 문서 쌍은 유사도 측정을 하지 않게 된다. 문서 벡터 생성 단계는 문서 클러스터링 결과와는 직접적인 관련이 없으므로 문서 벡터 DB를 생성한 다음 임계치를 조정해 가면서 상황에 맞는 문서 클러스터링을 수행할 수 있다.

4. 실험 및 결과 분석

본 논문에서 개발된 자동 색인 시스템과 문서 클러스터링 엔진에 대한 실험은 한국일보에서 제공한 3개월치 신문기사를 이용하여 수행되었다. 실험 데이터에 대한 설명은 다음 표와 같다.

|표 2| 실험 데이터 정보

데이터	한국일보기사(1999년 9월 - 1999년 11월)
건수	23,450(건)
크기	39.309(Mbyte)

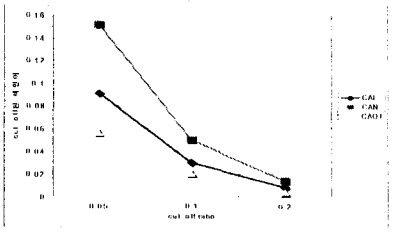
실험은 [표 1]에서 제시한 3가지 복합명사 색인 방법에 따라서 진행되었다. 각 유형별로 불용어 제거 임계치와 유사도 판정 임계치에 따른 클러스터링 결과의 변화를 관찰하였다. 우선 전체 데이터에 대해서 색인 작업을 수행하여 나온 유형별 색인어의 개수와 전체 색인어 개수와 제거(cut-off)된 불용어의 개수의 비율을 아래 표와 그림에 나타내었다.

|표 3| 실험 데이터 분석 정보

항목	유형		
	CAI	CAN	CAOI
단일화된 색인어 개수	297,731	176,651	286,966
전체 색인어 개수	5,260,083	4,602,929	3,995,321
색인어당 평균 출현빈도	17.667	26.057	13.923

CAN 방법에서 단일화된 색인어 수가 감소함을 알 수 있다. 이는 복합명사가 단위명사로 분리되면서 같은 단위명사가 많이 생기기 때문이다.

[그림4]에서 Y축은 제거된 불용어의 개수를 각 유형별 단일화된 색인어 개수로 나누어서 백분율로 나타낸 값이다. 위 그림에서 보는 바와 같이 CAN 타입으로 색인을 수행하였을 때 제거율이 가장 높게 나타난다. 이는 복합명사에서 분리된 단위명사가 일반적인 성향을 다수 띄어서 DF가 높게 나타나는 명사들이 많음을 보여준다. CAOI 타입에서 제거율이 낮은 이유는 복합명사가 분리되지 않으므로 복합명사를 구성하는 단위명사들이 제거되지 않기 때문이다.

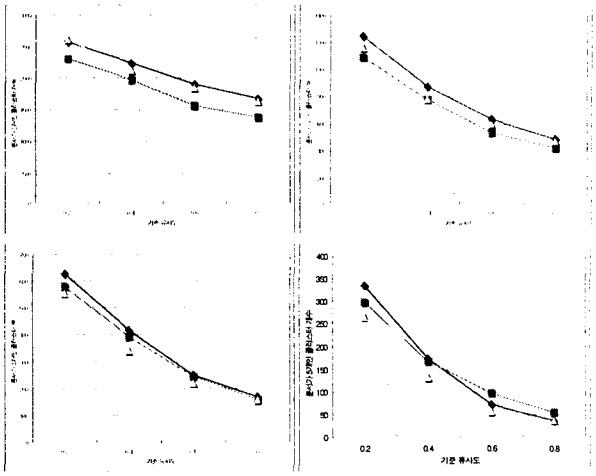


|그림 3| 색인어 Cut-Off 비율

|표 4| 유사도 임계치별 클러스터 개수

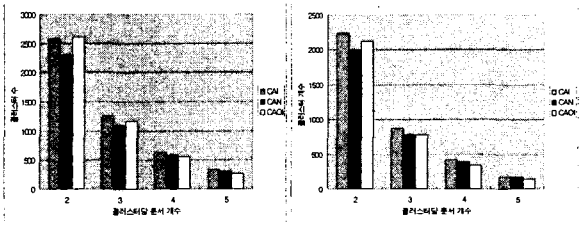
기준 유사도	CAI	CAN	CAOI
0.2	5,118	4,601	4,842
0.4	3,832	3,452	3,477
0.6	2,900	2,484	2,752
0.8	2,390	2,018	2,319
평균	3,560.00	3,138.75	3,347.50

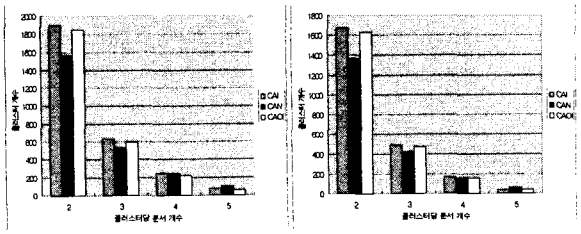
[표 4]는 유사도 임계치별 클러스터 개수를 보여준다. 평균적으로 CAI가 가장 많은 클러스터를 생성하고 CAN이 가장 적은 클러스터를 생성하게 된다.



|그림 5| 문서개수가 2,3,4,5인 클러스터 개수 변화 추이

[그림 5]는 각 유형별 기준 유사도에 따른 포함 문서가 2,3,4,5개인 클러스터 개수 변화 추이를 나타낸다. 문서개수가 증가함에 따라 CAOI의 분포가 낮아짐을 알 수 있다. 반면에 CAN은 문서개수가 증가함에 따라 클러스터의 개수 분포가 증가하고 있음을 보여준다. 각 유형별로 문서가 2개인 클러스터의 개수는 기준 유사도의 변화에 따른 값의 변화가 적다. 그러나 3개 이상인 클러스터의 개수는 기준 유사도의 변화에 따른 클러스터 개수의 변화가 큰 것을 알 수 있다.

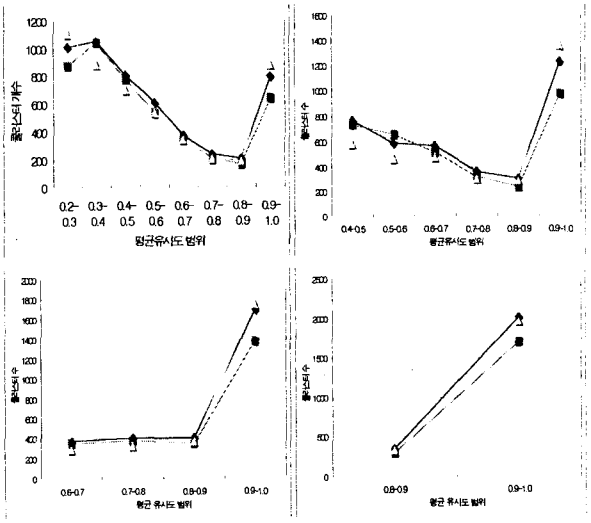




|그림 6| 기준 유사도별 클러스터당 문서개수 기준 클러스터 수

[그림 6]은 기준 유사도별로 클러스터 당 문서 개수를 기준으로 하여 클러스터의 개수를 각 유형별로 분석한 그래프이다. 매 기준 유사도에 따른 포함 문서가 2개인 클러스터의 개수 변화는 각 유형별로 변동이 적다. 그러나 포함 문서가 3개 이상인 클러스터의 개수 변화는 비교적 많이 드러남을 알 수 있다. 포함 문서 개수가 2개인 클러스터는 CAI와 CAOI 유형에서 많이 나타나지만 4개 이상인 클러스터는 CAN에서 비교적 많이 나타남을 보여준다.

[그림 7]은 기준 유사도별로 클러스터 내의 문서들 간의 평균 유사도 범위에 따른 클러스터의 개수를 나타낸다.



|그림 7| 평균 유사도 범위 기준 클러스터 개수 변화 추이

평균 유사도는 한 클러스터 내의 모든 문서 쌍에 대한 유사도 평균값이다. 이 값은 특정 클러스터의 응집도를 나타내기도 한다[6]. 평균 유사도 범위가 0.2-0.5인 곳을 제외하고는 세 가지 유형의 값 변화가 같은 형태로 나타난다. 또한 만일 기준 유사도와 클러스터 내의 평균 유사도 사이의 관계가 전혀 없다면, 위 네 개의 그래프는 동일한 형태를 나타내야 하나, 조금씩 차이를 나타내고 있는 것을 보면 HAC 알고리즘과 기준 유사도, 그리고 이를 통해 결과로 나오는 클러스터 집합의 평균 유사도 사이에는 어느 정도 연관성이 있음을 알 수 있다. 위 그림에서 CAOI 유형은 조금 특이한 형태를 취하고 있다. 평균 유사도 범위가 0.2-0.3인 경우를 제외하고는 0.4-0.5, 0.6-0.7 등의 범위 즉, 기준 유사도에 근접한 평균 유사도를 가지는 클러스터의 개수는 세 유형중 가장 낮지만, 가장 높은 평균 유사도인 0.9-1.0 사이의 클러스터는 가장 많이 나타난다. 또한 세 유형 모두 평균 유사도 0.8-0.9에서 클러스터의 수가 거의 동일하게 나타나고 있다.

5. 결론

본 논문에서는 구조화된 형태소 분석기를 이용하여 복합명사의 분석 유형을 다양하게 적용한 계층적 결합 문서 클러스터링 엔진을 개발하였다. 그리고 이를 통하여 결과로 나오는 클러스터 집합에 대한 실험 및 분석을 수행하였다. 클러스터링 결과에 대한 정확도는 테스트 집합의 부재로 실험하지 못하였으나, 본 기관의 내부 테스트 결과로는 CAI 유형의 클러스터링 집합이 사용자 만족도 면에서 가장 높은 점수를 획득하였다. 이는 일반적인 단위명사 분리의 잇점과 복합명사의 단일의미를 나타내는 복합명사 자체 색인어의 잇점이 부가되어서 나온 결과라고 생각된다.

복합명사 분석 방법과 클러스터링 알고리즘의 연관관계에 대한 보다 자세한 연구가 진행되어야 한다. 또한 문서 클러스터링 엔진에 대한 정확한 성능 측정을 위해서 문서 클러스터링 테스트 집합이 마련되어야 한다. 본 논문에서 수행된 실험 결과에 대한 정확한 수치적 의미를 부여하고 이를 통한 보다 세부적인 연구도 필요하다.

6. 참고문헌

[1] 최성필, 강무영, 주원균, 서정현, 김현, "자동 색인을 위한 한국어 형태소 분석기의 실제적인 구현 및 관리", KISTI Workshop 2001, 2001

[2] 최성필, "오류분석정보와 복합명사의 의미처리규칙 및 말뭉치를 이용한 철자 교정기의 성능개선", 부산대학교 전자계산학과 석사학위논문, 1998

[3] 심철민, "어절간 연관관계와 오류 유형 추정 규칙에 기반한 한국어 철자교정기", 부산대학교 전자계산학과 석사학위 논문, 1995

[4] 채영숙, 김재원, 김민정, 권혁철, "한국어 철자 검색을 위한 형태소 분석 기법", '91 우리말 정보화 잔치, 국어정보학회, pp.179-186

[5] 강승식, "음절 정보와 복수어 단위정보를 이용한 한국어 형태소 분석", 서울대학교 컴퓨터공학과 박사학위논문, 1993

[6] 강동혁, 주길홍, 이원석, "의미정보의 효율적인 분류를 위한 계층적 중복 문서 클러스터링", 한국정보과학회 2001 가을 학술발표논문집, 2001

[7] Manning, Christopher D. and Hinrich Schutze, "Foundations of Statistical Natural Language Processing", Cambridge: The MIT Press, 1999

[8] Baeza-Yates, Ricardo, and Ribeiro-Neto, Berthier, "Mordern Information Retrieval", New York: ACM Press, 1999

[9] Charniak, Eugene, "Statistical Language Learning", A Bradford Book, Cambridge: The MIT Press, 1993

[10] Sheldon Ross, "A First Course in Probability", Prentice Hall, 2002

[11] Ian H. Witten, Alistair Moffat, Timothy C. Bell, "Managing Gigabytes", Van Nostrand Reinhold, 1994

[12] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, "DATA MINING Methods for Knowledge Discovery", Kluwer Academic Publishers, 2000