개체명을 이용한 주제기반 웹 문서 클러스터링

Topic based Web Document Clustering using Named Entities

성기윤*, 윤보현**
이니텍(주)*, 목원대학교**

Ki-Youn Sung(yannnn@gmail.com)*, Bo-Hyun Yun(ybh@mokwon.ac.kr)**

요약

종래의 클러스터링 기법은 단순히 키워드를 추출에 기반한 단어간 유사도에 의한 그룹핑 방식을 구사함으로써 비교해야 할 대상 키워드 수 및 종류가 매우 다양하여 계산량이 증가함으로써 속도가 느리고 정확도도 높지 않은 편이다. 본 논문은 이러한 단점을 해소하기 위해 웹 문서를 대상으로 기존 명사 위주의 키워드 뿐 아니라 인명, 지명, 회사명, 물품명 등을 자동으로 인식하는 개체명 인식 결과를 이용하는 웹 클러스터링 기법을 제안하고자 한다. 실험을 통해 기존 키워드 기반 클러스터링 결과에 비해 개체명 기반 클러스터링의 품질이 우수함을 증명하였으며, 문서 집합 특성에 따른 클러스터링 결과도 비교 분석하였다.

■ 중심어 : | 클러스터링 | 개체명인식 | 자질추출 |

Abstract

Past clustering researches are focused on extraction of keyword for word similarity grouping. However, too many candidates to compare and compute bring high complexity, low speed and low accuracy. To overcome these weaknesses, this paper proposed a topical web document clustering model using not only keyword but also named entities such as person name, organization, location, and so on. By several experiments, we prove effects of our model compared with traditional model based on only keyword and analyze how different effects show according to characteristics of document collection.

■ keyword : | Creativity | Educational Game | Edutainment |

I. 서 론

일반적으로 검색에서 클러스터링(clustering)은 정보 접근성을 향상시키기 위한 정보 분류의 일종으로, 문서 분류(classification)이 학습 데이터를 기반으로 모델을 생성하는 교사학습 (supervised-learning)인 반면, 클러스터링은 입력값에 대한 목표치가 미리 주어지지 않는 비교사학습(unsupervised-learning) 이다. 주로 검색 결과 문서들을 실시간으로 분석하여 비슷한 주제의 문서들로 그룹을 나누어 분류한 결과를 제공해 관심 있는 주제를 중심으로 데이터를 조회할 수 있는 편리성을 제공하는데 활용된다.

최근 검색결과를 사용자에게 보다 지능적으로 재구성하여 제시하고자 하는 시도가 계속되고 있으며, 나아가 같은 주제를 다루고 있는 문서(예를 들면 신문기사)를 그룹핑하여 관리, 추적하는 와치 서비스(watch service)에 대한 요구가 대두 되고 있다.

기존의 클러스터링 기법은 단순히 키워드를 추출에

접수번호 : #100503-002
접수일자 : 2010년 05월 03일
심사완료일 : 2010년 05월 11일
교신저자 : 윤보현, e-mail : ybh@mokwon.ac.kr

기반한 키워드간 유사도에 의한 그룹핑 방식을 구사함으로써 비교해야 할 대상 키워드 수 및 종류가 매우 다양하여 계산량이 증가함으로써 속도가 느리고 정확도도 높지 않은 실정이다. 특히 대부분의 클러스터링 연구가 벡터 공간 모형 용어벡터로 표현하는 벡터공간모델에만 집중되어 다차원 개념 및 주제를 표현하기 어려운 단점이 있다.

이러한 문제점을 보안하기 위하여, 본 논문에서는 보다 정교한 언어분석을 통한 개체명 인식 결과를 바탕으로 인명, 지명, 회사명, 물품명 등 개념어 기반의 고차원 클러스터링 기법을 제안하고, 실험을 통해 기존 키워드 기반 클러스터링 결과에 비해 개체명 기반 클러스터링의 품질이 우수함을 증명하였다. 제안된 방법은 질의응답(Question Answering, [1])의 결과 제시시, 검색된 정답 문서와 답을 문서 클러스터링 결과로 제시할 때 활용될 예정이다.

논문의 구성은 다음과 같다. 2장에서는 기존 문서 클러스터링 기법에 대해 살펴보고, 3장에서는 본 논문에서 제안한 개체명을 이용한 주제기반 클러스터링 기법에 대해 설명하며, 4장에서는 이를 검증하기 위한 실험 및 결과에 대해 기술한다. 마지막으로 5장에서 결론을 맺도록 한다.

II. 관련 연구

기존 문서 클러스터링은 대부분 키워드 기반의 유사도에 의한 문서 그룹핑 방식으로 크게 계층적 클러스터링 (HAC: HierArchical Clustering) 기법[2]과 비계층적 클러스터링 기법[3][4]으로 나뉜다.

계층적 클러스터링은 비계층적인 방법에 비해 비교적 클러스터링 시간이 느리지만 매 단계에서 가장 유사한 문서 쌍을 우선으로 하여 단계적으로 클러스터 계층구조를 형성해 나가므로, 보다 정확한 클러스터링이 수행된다는 장점을 갖는다. 반면, 비계층적 클러스터링은 계층적 클러스터링에 비해 클러스터링 시간은 빠르지만 검색효율이 떨어지고 문서의 입력 순서에 따라 클러스터링 결과가 달라진다는 단점을 갖는다.

특히 질의응답을 위한 웹 문서 클러스터링의 경우 중요시 되는 것은 클러스터링의 속도보다는 클러스터링의 정확도이다. 따라서, 비계층적인 방법을 이용하기 보다는 속도는 다소 느리지만 정확도가 높은 계층적인 방법을 이용하는 것이 바람직하다. 클러스터링 알고리즘은 계층적인 클러스터링에서 가장 일반적으로 사용되는 세 가지 방법인 단일연결방법, 집단평균연결방법, 완전연결방법을 모두 이용하여 구현하였다.

특히, 기존의 클러스터링 기법은 단순히 키워드를 추출에 기반한 키워드간 유사도에 의한 그룹핑 방식을 구사함으로써 비교해야 할 대상 키워드 수 및 종류가 매우 다양하여 계산량이 증가함으로써 속도가 느리고 정확도도 높지 않은 실정이므로, 이를 보완하여 개체명을 중심으로 한 자질 추출을 통해 속도향상을 꾀한다.

개체명인식을 이용한 문서클러스터링에 관한 국외의 기술개발현황은 다음과 같이 크게 세가지 연구가 있다. Soto Montalvo [5]는 신문기사에서 다국어 뉴스 클러스터를 찾고자 개체명과 퍼지 유사도를 계산법을 이용한다. 문서 표현을 위해 개체명의 카테고리를 이용하고, 문서간의 유사도를 계산하기 위해 퍼지시스템에 기초한 접근법을 제안하고 있다. 실제 스페인과 영어 뉴스 문서집합에서의 실험을 통해 개체명이 다국어 뉴스 클러스터링을 위해 좋은 자질로 활용될 수 있음을 입증하였다.

Hiroyuki Toda [6]는 정보검색엔진의 검색결과를 대상으로 키워드 추출을 위해 개체명 인식을 사용하고, 검색결과에서 중요도를 판단하기 위해 레이블 선택 기준을 제시한다. 아울러 개체명 인식의 결과인 레이블 정보 이용한 레이블 범주화를 수행한다.

Wei, Gang [7]은 문서에 나타난 개체명을 통해 문서의 주제를 규정하였으며, 이를 바탕으로 문서 클러스터링을 수행하였다. 실험결과로 개체명 뿐 아니라 조응현상해소(co-referecne resolution)과 문맥의 활용이 필요하다고 밝히고 있다.

반면 외국의 경우에 비해 국내에서는 개체명 인식을 이용한 문서 클러스터링에 관한 연구는 전혀 없는 실정이다.

Ⅲ. 개체명을 이용한 주제기반 클러스터링

1. 개체명 기반 문서 벡터 생성

개체명 기반 문서 클러스터링을 수행하기 위해서는 각 문서를 위한 문서벡터를 생성해야 한다. 문서 벡터는 일반적으로 해당 문서를 구성하고 있는 단어들의 가중치(weight)를 원소로 갖는데, 본 논문에서는 그 대상 단어를 개체명(NE: Named Entity)과 일반 단어로 한정한다. 이때 개체명으로 인식된 단어는 최장단위로만 표현한다. 본 논문에서는 ETRI에서 개발된 개체명 인식기(NER: Named Entity Recognizer) [8]를 사용하였는데, 최상위에 사람(PERSON), 지역(LOCATION), 기관(ORGANIZATION) 등의 15 노드에 2~4 계층으로 구별된 총 147 노드로 구성된 개체명 태그를 제공한다.

각 문서를 위한 단어들의 가중치 계산 및 계산된 가중치들을 이용하여 구성되는 문서벡터는 다음의 [그림 2]와 같은 수식으로 표현될 수 있다.

$$weight_{ik} = \log(\frac{N}{df_k}) \times (\frac{tf_{ik}}{tf_{imax}}) \times boost_{N.E} \times Tag\_weight_{N.E}$$
$$D_i = (weight_{i1}, weight_{i2}, \cdots, weight_{i(L-1)}, weight_{iL})$$

그림 1. 단어 가중치 및 문서 백터 표현

위의 수식에서 $weight_{ik}$는 문서 i에서 단어(개체명+일반단어) k의 가중치를 나타낸다. $boost_{NE}$는 해당 단어가 개체명인가의 여부에 따라 부여되며, Tag_weight_{NE}는 해당 단어가 개체명인 경우, 각 개체명 태그별 (예: PERSON, LOCATION, ORGANIZATION 등) 가중치에 의미한다. N은 문서집합에 있는 전체 문서의 개수를 나타내며, df_k는 단어 k가 나타나는 문서의 개수를 나타낸다. tf_{ik}는 문서 i에서 단어 k가 나타나는 횟수이고, tf_{imax}는 문서 i에서 출현빈도가 가장 높은 단어의 출현빈도를 의미한다. 문서 i의 문서벡터 D_i는 문서집합 전체에서 유일한 단어의 개수 L에 대응하는 L차원 벡터로 표현되며, 각 원소는 각 단어의 가중치로 표현된다.

2. 개체명 기반 문서간 유사도 계산

단순 키워드 뿐만 아니라 개체명을 포함한 다양한 문서 자질 추출하고, 개체명을 반영하는 문서유사도 계산 방법 도출한다. 유사도를 계산하기 위해서 다음의 [그림 2]과 같은 다양한 유사계수(similarity coefficient)들을 사용할 수 있다. 본 연구에서는 [그림 2]에 있는 모든 방법을 이용할 수 있게 구현하였으며, Dice 유사계수를 이용한 방법을 기본으로 설정하고 있다[9][10].

Dice 유사계수(coefficient)

$$S_{D_i D_j} = \frac{2\sum_{k=1}^{L}(weight_{ik} weight_{jk})}{\sum_{k=1}^{L} weight_{ik}^2 + \sum_{k=1}^{L} weight_{jk}^2}$$

Jaccard 유사계수(coefficient)

$$S_{D_i D_j} = \frac{\sum_{k=1}^{L}(weight_{ik} weight_{jk})}{\sum_{k=1}^{L} weight_{ik}^2 + \sum_{k=1}^{L} weight_{jk}^2 - \sum_{k=1}^{L}(weight_{ik} weight_{jk})}$$

Cosine 유사계수(coefficient)

$$S_{D_i D_j} = \frac{2\sum_{k=1}^{L}(weight_{ik} weight_{jk})}{\sqrt{\sum_{k=1}^{L} weight_{ik}^2 + \sum_{k=1}^{L} weight_{jk}^2}}$$

그림 2. 유사 계수

S_{D_i,D_j} 는 문서 i와 j의 유사도를 나타낸다. 모든 문서 쌍에 대한 유사도가 계산되며, 계산된 유사도는 유사도 행렬(similarity matrix)에 저장된다. 또한 클러스터링의 기준으로 사용되는 유사도 값에 따라 문서 쌍을 정렬하기 위해 이진정렬트리(binary sort tree)를 구성한다.

이진정렬트리를 이용하면 트리의 루트로부터 최대 $\log_2 N$ 번의 비교를 거치면, 데이터를 그 크기에 따라 이진트리 구조상에 정렬할 수 있게 된다. 따라서, N개의 원소를 정렬하는 복잡도는 $O(N \log_2 N)$ 이 된다. 정렬된 데이터를 가장 유사도가 높은 문서 쌍을 우선으로 하여 순차적으로 이용하기 위해서는 이진정렬트리를 루트로부터 중위순회(inorder traversal)한다. 이를 계층적으로 클러스터를 구성해 나가는 과정과 결합하여 사용하면, 유사도가 높은 문서 쌍을 우선으로 추출할 수 있도록 해주기 때문에 전체적인 계층구조 구성에 소요되는 시간을 단축시킬 수 있다.

3. 개체명을 이용한 클러스터링

계층적 클러스터링은 아이템(item)이나 클러스터들의 쌍으로 이루어지는 중첩된 데이터 집합을 만들어내는 방법으로써, 이진트리(binary tree)와 유사한 형태의 클러스터 구조를 생성한다. 클러스터의 각 계층에서 반복적인 연산이 수행되어야 하므로 클러스터링 속도가 다소 느리지만, 클러스터링 결과에 대한 서로 다른 레벨의 해상도(resolution)를 제공할 수 있다는 장점을 갖는다. 다음의 [그림 3]은 계층적 클러스터링의 기본이 되는 방법론을 제시하고 있는 계층적 응집 클러스터링(Hierarchical agglomerative clustering: HAC)의 의사코드(pseudo code)이다.

```
Initialize all documents as singleton cluster
Until (Halting criterion) do {
    Find two most similar clusters
    Merge them
}
```

그림 3. 계층적 응집 클러스터 의사코드

문서집합에 있는 각 문서를 하나의 클러스터로 가정한 상태에서 시작하여, 각 단계에서 반복적으로 가장 유사한 클러스터 쌍을 병합하면서 클러스터를 만들어내는 방법으로써, 클러스터간의 거리함수(distance function)와 정지 기준에 따라 클러스터링 결과가 좌우된다.

본 논문에서 정지기준은 문서 쌍들에 대해 계산된 유사도 중에서 더 이상 유효한 유사도가 없거나, 모든 문서들이 클러스터 계층구조에 포함되었을 때로 주어졌다. 유효한 유사도가 없다는 것의 의미는 유사도에 따라 정렬된 문서 쌍들에서 그 값이 0보다 큰 것이 더 이상 없다는 것을 의미한다.

계층적 클러스터링에서 클러스터간의 거리함수로 사용되는 것은 2장에서 설명한 바와 같이, 단일연결(single-link)함수, 완전연결(complete-link)함수, 집단평균연결(group-average-link)함수[1]의 세 가지가 있다.

본 논문에서는 계산된 유사도 행렬과 이진정렬트리를 이용하여 세 가지 거리함수 즉, 단일연결, 완전연결, 집단평균연결함수 모두를 구현하였다. 클러스터 계층구조는 상향식(bottom-up)으로 구성되며, 계층구조의 중간노드에 해당되는 모든 노드들은 각기 하나의 클러스터를 나타낸다. 계층구조의 종단노드에 해당되는 모든 노드는 문서집합에 있는 각 문서를 나타냄과 동시에 최소단위 클러스터(singleton cluster)를 나타낸다.

Ⅳ. 실험 및 평가

본 논문에서는 구현된 개체명 기반 클러스터링 시스템의 성능을 알아보기 위해 수작업으로 분류된 문서집합을 입력으로 주어서 클러스터링을 수행하였다. 클러스터링 시스템의 성능은 클러스터링 결과로 생성된 문서그룹이 얼마나 잘 응집되어 있는지에 따라 결정된다. 문서그룹의 응집성이 높다는 것은 하나의 문서그룹이 다른 문서그룹과 뚜렷하게 구분되면서, 내부에 나타난 모든 문서가 강한 연관성을 가진다는 것을 의미한다. 이를 명확하게 판단하기 위해서 사람이 직접 문서를 분석하여 분류한 문서집합을 이용하여 클러스터링을 테

스트하였다.

다음의 [표 1]과 [표 2]에 기술된 두 개의 테스트 집합은 개체명 기반 클러스터링시스템의 성능 평가를 위해 수작업으로 미리 분류한 문서집합이다. 첫 번째 문서집합은 각 범주의 개념이 의미적으로 비교적 큰 차이를 갖고 있는 반면에, 두 번째 문서집합은 “어음”과 “보험”이라는 상위 개념 하에 나타날 수 있는 유사한 개념들을 각 범주로 나누어 표현하고 있다.

표 1. 상품 리뷰 문서 집합 (문서 수: 50, 범주 수: 50)

자동차	금융	생활가전/ 통신기기	쇼핑/ 택배	스포츠/ 레저용품
1~10	11~20	21~30	31~40	41~50

표 2. 경제 문서 집합 (문서 수: 247, 범주 수: 8)

견질어음	1~11	융통어음	89~94
기한부어음	12~23	고용보험	95~146
백지어음	24~44	보상보험	147~195
약속어음	45~88	의료보험	196~247

[그림 4]은 상품 리뷰 문서 중 “자동차” 관련 문서에 대한 언어분석 결과를 나타낸다. 밑줄로 표시된 부분은 개체명으로 인식된 예를, "i30"은 개체명이나 언어분석 오류로 인해 일반 명사로 분석된 예를 나타낸다. [그림 5]는 해당 문서에서 추출된 자질로, 자질 가중치 별로 정렬된 예이다.

클러스터링 테스트 결과는 다음과 같은 평가 요소들을 이용하여 표현되었다.

- A : 단일범주 클러스터 노드에서의 평균 문서 수
- B : 클러스터 평균 정확도

평가지표 A는 단일범주 문서로 이루어진 노드들이 포함하고 있는 문서 개수의 평균값으로, 평가셋 대비 그룹핑된 겨로가 클러스터가 쏠림현상이 있는지를 알 수 있는 척도가 된다. 평가지표 B는 현재 레벨에 존재하는 모든 노드들에서 (클러스터 내에서 주류를 이루는 범주에 속하는 문서 개수/클러스터내 전체 문서 수) 값을 구하여 이를 평균 낸 값을 의미한다. 이 값이 1에 가

타봤습니다! / <현대자동차:ORG_BUSINESS> <i30:AF_TRANSPORT>

‘나는 다르다. 그래서 선택한 차가 이 차’라는 광고 카피를 앞세운 <i30:AF_TRANSPORT>은 <현대자동차:ORG_BUSINESS>의 야심작이다.

<아반떼::AF_TRANSPORT>와 동급인 배기량 1.6L이지만 승차감은 차이가 많이 났다. ….. 브레이크도 반응 속도가 좀 느린 다른 <현대차::ORG_BUSINESS>와는 달리 빠르게 반응했다. 뒤에서 따라붙는 차의 제동력을 걱정해야 할 정도였다…. <1000만원:QT_COUNT>대의 해치백이라는 점을 고려한다면 <3000만원:QT_COUNT>대인 <골프:AF_TRANSPORT>나 <C30:AF_TRANSPORT> 등과 가격 대비 경쟁에서 우월하다고 평가할 만했다.

플랫폼(차 뼈대)을 공유하고 있는 <기아차::ORG_BUSINESS>의 ‘씨드’와 차이는 크지 않았다. 옆모습만 봐서는 구별이 힘들다. 앞에서 보면 i30은 범퍼그릴을, 씨드는 라디에이터그릴을 더 강조한 차이가 있다. 씨드는 실내조명이 붉은색인 반면 i30는 푸른색이다. i30은 씨드에 비해 서스펜션이 단단하지는 않은 느낌이다. “안락한 승차감을 좋아하는 국내 운전자들의 취향을 어느 정도 고려했기 때문”이라는 게 <현대차::ORG_BUSINESS> 관계자의 설명이다.

그림 4. 언어분석 결과 예 (개체명 인식 결과)

```
( 현대자동차, 0.356499124318, OGG_BUSINESS, 14 100 )
( i30, 0.25263124318, AF_TRANSPORT, 24 92 )
( 현대차, 0.138720827177, OGG_BUSINESS, 728 1702 )
( 기아차, 0.10666693002, OGG_BUSINESS, 1364 )
( 스포츠카, 0.03880248731, AF_TRANSPORT, 1035 )
( 문병주, 0.0346269570291, PS_NAME, 1730 )
( 아반떼, 0.0290759140626, AF_TRANSPORT 423 484 )
( 씨드, 0.028915042989, AF_TRANSPORT 1375 1465 1512 1568 )
( 차량, 0.0241320785135, np, 1077 )
( 가속, 0.0239303745329, np, 976 1120 )
( 해치백, 0.0216573290527, np, 240 1163 )
```

그림 5. 자질 추출 예

까울수록 현재 레벨에 나타난 클러스터들의 순수도가 높다. 설명에서 사용되는 "레벨(level)"은 클러스터 계층 구조 상에서의 루트(root)로부터 시작되는 트리의 깊이(depth)를 나타낸다.

표 3. 상품 리뷰 문서 집합 테스트 결과

	Single Link		Complete Link		Group Average Link	
	A	B	A	B	A	B
1	–	0.2	–	0.2	–	0.2
2	–	0.442	–	0.442	2	0.604
3	6	0.773	6	0.773	4	0.815
4	6.333	0.906	***4.857***	***0.953***	5	0.694
5	3.75	1	3.8181	0.972	2	0.911
6	2.5	1	2.222	0.962	2.25	0.853
7	***6.333***	***0.916***	1.68	1	5.4	0.935
8	2.5	1	1.571	1	***3.09***	***0.989***
9	2.833	1	1.66	1	2.25	1
10	2	0.906			2.166	1
11	2	1			1.666	1
12	1.416	1			1.5	1
13	1.75	1			1	1
14	1.5	1				
15	1	1				
16	1	1				

* 해당 클러스터링 기법 중 최적의 결과

[표 3]는 상품 리뷰 문서집합을 입력으로 하여 개체명 기반 클러스터링을 수행한 결과를 정리한 것이고, [표 4]는 경제 문서집합을 입력으로 하였을 경우의 결과이다. 각각은 클러스터링 평가 요소를 이용하여 클러스터 계층구조의 깊이에 따른 클러스터링 결과를 나타내고 있다.

[표 3]와 [표 4] 모두에서 단일연결(single link) 방법을 이용한 경우 계층구조의 깊이가 다른 방법들에 비해 깊게 나타남을 알 수 있다. 이는 단일연결 방법이 전체적으로 경사진 이진트리 형태의 계층구조를 형성하기 때문에 일어나는 현상이다. [표 4]의 경우 깊이 40이상의 데이터를 생략하여 나타낸 결과인데, 실제 데이터는 단일연결이 104, 완전연결(complete link)이 41, 집단평균연결(group average link)이 59로 그 깊이에 있어 큰 차이를 보인다. 생성된 클러스터의 질은 "클러스터 평균 정확도"를 봄으로써 판단할 수 있다. 각 레벨에서 계산된 "클러스터 평균 정확도"는 해당 레벨에 나타나는 모든 클러스터 노드들에 얼마만큼 순수하게 한 가지 범주만이 포함되어 있는지에 대한 정도를 파악할 수 있는 기준이 된다.

표 4. 경제 문서 집합 테스트 결과

	Single Link		Complete Link		Group Average Link	
	A	B	A	B	A	B
1	–	0.21	–	0.21	–	0.21
2	1	0.60	3	0.60	–	0.35
3	2	0.60	2	0.80	1	0.80
4	1	0.67	1.8	0.86	1	0.60
5	1	0.80	1.4	0.87	1	0.60
6	1	0.60	1	0.86	2	0.60
7	1	0.60	1	0.60	1.33	0.80
8	–	0.35	–	0.35	1	0.80
9	1.3	0.80	1	0.67	1	0.60
10	1	0.44	1	0.80	–	035
11	–	0.80	–	0.40	1	0.80
12	1	0.80	1	0.70	1	0.60
13	–	0.60	–	0.43	–	0.35
14	1	0.61	1	0.77	1	0.63
15	2	0.61	1.5	0.77	1.5	0.68
16	1	0.61	1.75	0.78	1.2	0.86
17	1	0.608	1.23	0.68	1	0.80
18	1	0.804	1.11	0.89	1	0.60
19	1	0.609	1.2	0.87	1	0.60
20	1	0.61	1.25	0.86	1	0.60
21	1	0.61	1.85	0.87	1	0.61
22	1	0.609	***6.42***	***0.915***	–	0.61
23	–	0.609	5.4	0.91	1	0.54
24	1	0.607	3.15	0.97	2.5	0.76
25	1	0.605	3.0	0.96	1.4	0.68
26	1	0.606	5.25	0.96	1.4	0.87
27	–	0.35	5	0.91	1.333	0.8
28	1	0.803	7.90	0.95	1	0.80
29	1	0.608	5.17	0.91	2	0.80
30	1	0.609	5.31	0.92	1	0.61
31	1	0.609	2.56	0.94	1	0.62
32	1	0.61	3.74	0.97	–	0.81
33	1	0.61	2.5	1	2	0.83
34	1	0.61	2.09	1	1.2	0.82
35	2	0.612	1.75	1	1	0,84
36	1	0.806	1.6	1	2	0.82
37	1	0.613	1.7	1	2.142	0.93
38	1	0.611	1.37	1	***3.41***	***0.95***
39	1	0.611	1.25	1	3.77	0.93
40	1	0.612	1	1	4.9	0.88

첫 번째 테스트 문서집합의 경우 정확도 0.9 이상이 나타나는 레벨이 단일연결의 경우 4, 완전연결의 경우

4, 집단평균연결의 경우 5로 비슷한 레벨에서 나타남을 알 수 있다. 그러나 두 번째 테스트 문서집합의 경우는 0.9 이상의 정확도를 갖는 레벨이 단일연결은 91, 완전연결은 22 집단평균연결은 37로 큰 차이를 보이고 있다. 이러한 실험 결과 계층적 클러스터링 방법을 이용한 클러스터링에 있어서 응집성이 있는 양질의 클러스터를 생성하기 위해서는 완전연결 방법을 이용하는 것이 가장 좋음을 알 수 있었다.

리뷰 문서집합의 경우 각 범주의 의미가 뚜렷이 구분되는 특성을 가지기 때문에 완전연결이나 집단평균연결 방법을 이용한 클러스터링의 결과가 실제 수작업으로 만들어진 범주와 대부분 일치하는 좋은 결과를 보이고 있다. 그러나, 경제 문서집합의 경우 “어음”과 “보험”이라는 큰 범주들끼리의 그룹핑은 대체적으로 잘 이루어지는 반면에 그 이하의 세부적인 범주로의 그룹핑은 잘 이루어지지 않았음을 알 수 있었다. 이는 테스트 문서집합에서 세부적인 범주들에 나타나는 문서들이 이웃 범주에 나타난 문서들과 뚜렷이 구분되는 특징적인 개체명 자질 집합으로 이루어져 있지 않고, 유사한 개체명 및 기존 단어 자질들을 많이 포함하고 있기 때문에 일어나는 현상이다. 따라서, 본 연구에서 이용한 중심 개체명의 빈도에 의존한 문서특성 파악 방법으로는 세부적인 범주에 적합한 클러스터링을 하기 어려운 문서집합이다.

다음 표는 기존 키워드 기반 클러스터링 결과와 본 연구에서 제안한 개체명 기반 클러스터링 결과를 비교한 것이다. 상품 리뷰 문서집합의 결과를 자세히 설명해 보면, 기존 키워드 기반 클러스터링 결과의 경우, 0.9 이상의 클러스터 레벨 (7)이 개체명 기반 결과 (4)에 비해 3단계 깊은 것을 볼 수 있다. 또한 단일 범주 문서로 이루어진 노드의 평균 문서 수 역시 키워드 기반은 3.2인 반면, 개체명 기반 클러스터링 결과에서는 4.857로 같은 군집의 문서들이 그룹핑이 많이 되었음을 알 수 있다. 이러한 결과는 경제 문서집합에서도 같은 결과를 보였는데, 키워드 기반 클러스터링 결과의 경우, 0.9 이상의 클러스터 레벨 (22)이 개체명 기반 결과 (55)에 비해 무려 30여 단계가 깊은 것을 볼 수 있다

표 5. 키워드 vs. 개체명 기반 클러스터링 결과 비교

		평균클러스터링 정확도 0.9 이상의 클러스터 레벨	단일범주 클러스터 노드에서의 평균 문서 수
상품 리뷰	키워드 기반	7	3.2
	개체명 기반	4	4.857
경제	키워드 기반	55	2.1
	개체명 기반	22	6.42

이는 기존 키워드 기반 자질 추출이 단어, 형태소 기반으로 구성되어 의미적 공통성을 발견하기 어려우며, 자료 빈곤(data sparseness)문제가 심각하게 발생하기 때문이다. 이에 비해 개체명 기반 클러스터링은 같은 개체명 태그 정보 관점에서 자질을 재구성하여 유사도를 계산하기 때문에 이러한 문제에 영향을 덜 받게 되며, 개체명 태그는 또한 문서의“주제”를 대표할 수 있는 후보로 활용 가능하다.

Ⅵ. 결론

본 논문에서는 본 연구는 웹 문서를 대상으로 키워드 뿐 아니라 인명, 지명, 회사명, 물품명 등을 자동으로 인식하는 개체명 인식 결과를 이용하는 웹 클러스터링 기법을 제안하였다. 실험을 통해 기존 키워드 기반 클러스터링 결과에 비해 개체명 기반 클러스터링의 품질이 우수함을 증명하였다.

그러나 개체명 기반 클러스터링에서도 의미적(Semantic level) 자질 분석시 같은 의미이지만 다른 형태로 표현된 경우, 예를 들면 “현대차, 현대자동차, 현대”가 모두 같은 회사명을 나타내는 표현임을 파악하지 못해 발생한 문제는 여전히 남아있다.

또한 클러스터링 알고리즘 상의 과쏠림 현상이 발생하였는데, 문서 유사도 기반이나 전체 문서집합을 대상으로 한 것이 아닌, 수집된 문서(or 검색 결과)를 대상으로 한 것이 아니므로 역문서빈도(idf) 값이 왜곡되기 때문이며, 특히 개체명 태그 위주로 쏠림 현상이 발생하였다. 향후에는 이러한 문제점을 해결하기 위한 연구를 진행할 예정이다.

참 고 문 헌

[1] H. J. Oh, S. H. Myaeng, and M. G. Jang, "Enhancing Performance with a Learnable Strategy for Multiple Question Answering Modules," ETRI Journal, Vol.31, No.4, 2009.

[2] Oren Zamir, "Fast and Intuitive Clustering of Web Documents," Qual's Paper, University of Washington.

[3] Oren Zamir and Oren Etzioni, "Web Document Clustering: A Feasibility Demonstration," Proc. of ACM SIGIR'98, 1998.

[4] Oren Zamir and Oren Etzioni, "Grouper: A Dynamic Clustering Interface to Web Search Results," Proc. of WWW8, pp.1361-1374, 2009.

[5] Soto Montalvo and Raquel Martinex, "Bilingual New Clustering Using Named Entities and Fuzzy Similarity," Proc. of 10th TSD, 2007.

[6] Hiroyuki Toda and Ryoji Kataoka, "search result clustering method using informatively named entities," Proc. of ACM internationa workshop on WIDM, pp.1-86, 2005.

[7] Gang Wei, "Named Entity Recognition and An Apply on Document Clustering," MSCs thesis, Dalhousie University, 2004.

[8] C. K. Lee, Y. G. Hwang, and S. J. Lim, "Fine-Grained Named Entity Recognition Using Conditional Random Fields for Question Answering," Proc. AIRS-06, LNCS Vol.4182, pp.581-587, 2006.

[9] B. William, Frakes, and Richard Baeza-Yates, "Clustering Algorithm," Information Retrieval Data Structure and Algorithm, Chapter 16.

[10] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, "Modern Information Retrieval," Addison-Wesley, 1999.

저 자 소 개

성 기 윤(Ki-Youn Sung) 정회원

- 1995년 : 충남대학교 컴퓨터과학과(학사)
- 2003년 : 충남대학교 컴퓨터과학과(석사)
- 1995년 ~ 2000년 : (주)우린정보 대리
- 2000년 ~ 2001년 : (주)예스월드 팀장
- 2001년 ~ 2002년 : (주)루틸러스 팀장
- 2002년 ~ 현재 : 이니텍(주) 부장

<관심 분야> : 정보검색, 인터넷 뱅킹, 미들웨어, 지식처리, 스마트 서비스

윤 보 현(Bo-Hyun Yun) 정회원

- 1992년 : 목포대학교 전산통계학과(공학사)
- 1995년 : 고려대학교 컴퓨터학과(이학석사)
- 1999년 : 고려대학교 컴퓨터학과(이학박사)
- 1999년 ~ 2003년 : 한국전자통신연구원 팀장
- 2003년 ~ 현재 : 목원대학교 컴퓨터교육과 교수

<관심분야> : 자연어처리, 정보검색, Semantic Web