# 온톨로지를 이용한 단어 군집화 성능 개선

박 은 진[†] · 김 재 훈[††] · 옥 철 영[†††]


## 요 약

이 논문은 사전의 뜻 풀이말을 이용하여 단어 군집화 시스템을 설계하고 구현한다. 군집화를 위해서는 다양한 형태의 자질이 요구되며 어떤 자질을 사용하느냐에 따라 군집화의 성능이 좌우된다. 뜻 풀이말은 표제어를 자세히 설명하고 있기는 하지만, 뜻 풀이말에 사용된 단어가 너무 함축적이거나 추상적이어서 뜻 풀이말이 그다지 길지 않다. 뜻 풀이말로부터 추출된 자질을 그대로 군집화에 이용할 경우에는 다수의 작은 군집이 형성된다. 뜻 풀이말을 이용하여 보다 더 좋은 군집화 결과를 얻기 위해서는 뜻 풀이말의 의미를 크게 손상하지 않는 범위에서 보다 더 일반적인 단어로 바꾸어 군집화에 필요한 자질을 확장할 필요가 있다. 이 논문에서 추상적인 말을 온톨로지 상에서 한 단계 위의 단어로 확장하거나 온톨로지 상에서 고정 높이에 해당하는 단어로 확장함으로써 단어 군집화 성능을 향상시키는 방법을 제안한다. 실험을 통해서 온톨로지를 이용해서 자질을 확장할 경우 단어 군집화 성능이 크게 개선되었으며, 전체적으로 보면 온톨로지 상에서 고정 높이에 해당하는 단어로 확장할 경우가 더 좋은 성능을 보였다. 또한 단어 군집화를 위한 자질로 동사가 매우 유용함을 관찰할 수 있었다.


**키워드 : 온톨로지, 단어 군집화, 자질 확장**

# Performance Improvement of Word Clustering Using Ontology

Eun-Jin Park[†] · Jae-Hoon Kim[††] · Cheol-Young Ock[†††]


## ABSTRACT

In this paper, we describe the design and the implementation of word clustering system using a definition of an entry word in the dictionary, called a dictionary definition. Generally word clustering needs various features like words and the performance of a system for the word clustering depends on using some kinds of features. Dictionary definition describes the meaning of an entry in detail, but words in the dictionary definition are implicative or abstractive, and then its length is not long. The word clustering using only features extracted from the dictionary definition results in a lots of small-size clusters. In order to make large-size clusters and improve the performance, we need to transform the features into more general words with keeping the original meaning of the dictionary definition as intact as possible. In this paper, we propose two methods for extending the dictionary definition using ontology. One is to extend the dictionary definition to parent words on the ontology and the other is to extend the dictionary definition to some words in fixed depth from the root of the ontology. Through our experiments, we have observed that the proposed systems outperform that without extending features, and the latter's extending method overtakes the former's extending method in performance. We have also observed that verbs are very useful in extending features in the case of word clustering.


**Key Words : Ontology, Word Clustering, Feature Extension**

## 1. 서 론

인터넷의 대중화와 양적인 팽창으로 온라인 상에서 획득할 수 있는 정보의 양이 급격히 증가하고 있다. 정보의 더미 속에서 양질의 정보를 찾아내는 것은 대단히 어려운 문제이다. 이를 위해 Naver[1], Google[2] 등과 같은 인터넷 정보 검색 엔진들이 널리 사용되고 있으나 일반적으로 검색 결과의 문서 수가 너무 많아서 원하는 정보를 찾기는 여전히 어렵다. 최근 이러한 문제를 해결하기 위하여, 방대한 정보를 스스로 기계가 학습하여 연관된 정보를 한 곳에 모으는 기법인 군집화(clustering)에 관한 연구가 활발히 이루어지고 있다[1-3].

군집화는 군집 대상이 텍스트일 경우에 문서 군집화(document clustering)와 단어 군집화(word clustering)로 분

※ 이 논문은 2004년 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2004-002-D00372).
† 정 회 원 : 한국해양대학교 컴퓨터공학과 석사과정
†† 종신회원 : 한국해양대학교 컴퓨터공학과 부교수
††† 종신회원 : 울산대학교 컴퓨터정보통신공학부 교수
논문접수 : 2006년 3월 15일, 심사완료 : 2006년 4월 12일

1) http://www.naver.com/
2) http://www.google.co.kr/

류할 수 있다. 문서 군집화는 검색 엔진의 검색 결과의 군집화[1-3], 문서 자동 요약[4], 사건 탐색 및 추적[5] 등에서 사용되고 있으며, 단어 군집화는 용어의 모호성 해소[6], 정보 검색 시스템의 질의 확장[7] 등에 사용되고 있다.

이 논문은 사전의 뜻 풀이말을 이용한 단어 군집화의 성능향상에 관련된다. 일반적으로 사전의 뜻 풀이말은 함축적이고 추상적인 말로 표제어를 설명한다. 예를 들어 '강아지'는 "개의 새끼"로 설명되어 있다. 여기서 '개'는 표제어 '강아지'를 포함하는 추상적인 말이며 온톨로지 상에서 '강아지'보다 위에 있다. 바꿔서 말하면 뜻 풀이말에 쓰인 '개'는 '강아지'와 같은 군집이 될 수 있다. 이러한 특징으로 인하여 단어를 군집할 때, 뜻 풀이말을 단어 군집화의 자질로 사용하면 양질의 군집을 형성할 수 있을 것이다.

그러나 사전의 뜻 풀이말 자체는 너무 함축적으로 단어를 표현하기 때문에 자질의 수가 매우 작다. 이러한 특징은 뜻 풀이말을 이용한 단어 군집화 결과가 다수의 작은 군집으로 나타난다. 이 논문에서는 다수의 작은 군집을 양질의 큰 군집으로 만들기 위하여 사전의 뜻 풀이말로 사용된 단어(혹은 자질)를 온톨로지 상에서 한 단계 위의 단어로 확장하거나 최상위 개념에서 특정한 높이에 있는 단어로 확장함으로써 단어 군집화 성능을 향상시키는 방법을 제안한다.

이 논문은 다음과 같이 구성된다. 2장에서는 관련 연구를 살펴보고 3장에서는 단어 군집화 시스템의 구성을 설명한다. 4장에서는 자질의 확장에 따른 단어 군집화 성능을 비교하고 분석한다. 마지막으로 5장에서는 실험 결과를 바탕으로 결론을 맺고 향후 연구 과제를 설명한다.

## 2. 관련 연구

이 장에서는 단어 군집화에 관한 연구, 사전의 뜻 풀이말을 이용하는 연구, 온톨로지, 군집화 평가에 관한 연구를 간략히 살펴본다.

### 2.1 단어 군집화

단어 군집화는 사람의 개입 없이 기계가 스스로 학습하여 의미가 유사한 단어를 하나로 모으는 자율 기계학습 방법이다. 단어 군집화에서 필요한 자질을 추출하는 방법으로 단어의 용례를 자질로 이용한 방법[8]과 사전의 뜻 풀이말을 자질로 이용한 방법[9]으로 분류할 수 있다. 단어의 용례를 자질로 이용하는 방법은 "유사한 단어는 비슷한 용례를 가진다"라는 가정에 근거한다. 이 방법은 대규모 말뭉치에서 단어 bigram 혹은 격 정보를 추출하여 군집화를 위한 자질로 사용한다. 이 방법은 상대적으로 사용 빈도가 낮은 단어에 대해서는 출현 빈도가 낮기 때문에 자질 벡터 형성이 어렵고 연관성이 있는 단어라 할지라도 같이 나오는 확률이 낮을 수 있기 때문에 같은 군집을 형성하기 어렵다. 단어의 뜻 풀이말을 자질로 이용하는 방법은 "유사한 단어는 비슷한 뜻 풀이말을 가진다"라는 가정에 근거한다. 이 연구에서는 사전의 뜻 풀이말과 온톨로지에서 표제어의 위치 정보를 이용하여 단어 군집을 형성하고 이를 바탕으로 단어의 모호성을 해소하고 정보 검색에 적용했다[9].

단어 군집화는 유사도 계산 방법에 따라서도 의미 유사도(semantic similarity)와 의미 관련도(semantic relatedness)로 분류될 수 있다[10]. 의미 유사도는 온톨로지 상에서 상위어와 동의어 관계를 이용하거나(taxonomy-based semantic similarity)[9] 단어의 분포 관계(distributional evidence)를 이용하여(distributionally-based semantic similarity)[11] 유사도를 계산하는 방법이고, 의미 관련도는 자연언어처리 문제를 해결하는데 있어서 관련 정도를 이용하는 방법이다[12].

### 2.2 사전의 뜻 풀이말을 이용하는 연구

사전 뜻 풀이말의 중첩된 정도를 이용하여 단어의 모호성을 제거하는 연구가 있어 왔다[9, 13]. 이 연구에서는 뜻 풀이말에 같은 말이 나타나면 서로 연관이 있는 단어로 가정한다. 그러나 이 연구에서는 사전 뜻 풀이말이 작아서 서로 겹치지 않는 것이 문제가 된다. 이러한 문제를 해결하려는 연구가 있었다[14]. 이 연구에서는 사전의 뜻 풀이말을 대량의 말뭉치에서 추출한 언어정보로 확장하여 뜻 풀이말의 크기를 확장하였다. 그러나 이 방법은 말뭉치에 따라 빈번히 사용되는 단어가 다르기 때문에 성능이 다르고 상대적으로 사용빈도가 낮은 단어에 대해서는 자질 확장이 어려운 단점이 있다.

사전의 뜻 풀이말 특성을 이용해 정제한 의미정보를 확률정보, 거리정보 및 격정보, 문장 분할정보 등을 기반으로 단어의 의미 모호성을 해결하려는 연구가 있어 왔다[15]. 이 연구에서는 뜻 풀이말 크기 문제를 해결하기 위해 울산대학교의 UWIN[16, 17]의 단어 계층적 구조를 이용하였다. 사전의 뜻 풀이말에서 계층 정보를 추출하여 어휘 계층망을 구축하려는 연구가 있어 왔다[18]. 이 연구에서는 사전의 뜻 풀이말의 마지막 풀이말이 계층 정보를 포함한다고 가정하고 이를 단어의 의미 계층 형성에 이용하였다.

### 2.3 온톨로지

국내에서 개발중인 온톨로지로는 카이스트의 CoreNet[19]을 비롯하여 오롬 정보의 NexusBase[20], 부산대학교의 KorLex[21], 전자통신연구원의 ETRI 어휘개념망[17], 울산대학교의 UWIN[16, 17] 등이 있다. CoreNet의 경우 개념기반의 다국어 어휘의미망으로써 한국어, 중국어, 일본어로 구축되어 있고 단일어 사전과 기존의 워드넷을 이용하여 반자동으로 구축되었으며 자연언어처리 및 의미기반 지식처리 시스템에 활용하고 있다. NexusBase는 국제 표준에 맞추어 구축 중인 국내 최대 규모로 40만 용어 이상을 포함하고 있고 오롬 시소러스 시스템과 연동되어 있으며 다국어 시소러스 형태로 구축 중이다. KorLex는 한국형 워드넷(WordNet[22])으로써 워드넷의 한국어 번역 결과이다. ETRI 어휘개념망은 현재 개체명 사전까지 연결된 개념망으로 백과사전 기반 질의응답 시스템에 활용된다. UWIN은 단어의 사전적

뜻 풀이말을 바탕으로 단어의 세부의미 수준까지 계층 분류가 되어 있고 단어의 모호성 해소 시스템 및 형태소 분석기 등에 응용되고 있다. 이 논문에서는 결과 군집 평가와 자질을 확장할 때 울산대학교의 UWIN을 사용하였다.

### 2.4 군집화 평가 방법

군집화 평가는 방법에 따라 크게 세 가지로 분류된다[23, 24]. 사람이 미리 정의한 군집과 기계가 형성한 군집을 비교하는 방법인 **외부 평가**(external criteria)와 기계가 형성한 군집 결과를 평가할 때 외부 기준을 사용하지 않고 데이터 자체만을 이용하여 평가하는 방법인 **내부 평가**(internal criteria), 그리고 통계적인 방법을 이용하지 않고 동일한 군집화 알고리즘의 최상의 군집 환경(자질 표현 및 유사도 측정)을 찾는 방법인 **상대 평가**(relative criteria)가 있다. 각각의 대표적인 평가 방법은 <표 1>과 같다[23, 25].

〈표 1〉 군집화 평가 방법

| | |
|---|---|
| 외부 평가 | *Rand Statistic, Jaccard Coefficient, Folks and Mallows Index, Huberts Γ Statistic* |
| 내부 평가 | *Cophenetic Correlation Coefficient, Hubert's Γ Statistic* |
| 상대 평가 | *Modified Hubert Γ Statistic, Dunn Index, Davies-Bouldin Index* |

이 논문에서는 외부 평가(*Rand Statistic, Jaccard Coefficient, Folkes and Mallows Index, F-measure*)와 상대 평가(*Dunn Indices, Davies-Bouldin Index*)를 이용하여 입력 자질에 따른 단어 군집화 성능을 평가할 것이다.

#### 2.4.1 외부 평가

외부 평가는 평가를 위해서 미리 정의해둔 정답 군집을 이용하며, 주로 <표 2>와 같은 이원 분할표를 활용한다. 이 논문에서는 *Rand Statistic, Jaccard Coefficient, Folkes and Mallows Index, F-measure*을 이용할 것이다.

결과 군집의 개수와 정답 군집의 개수가 여러 개이므로 이원 분할표를 계산하기 위해 모든 입력 단어의 단어 쌍 $(x_i, x_j)$를 만든다. <표 2>에서 a는 임의의 단어 쌍이 같은 정답 군집에 있고 같은 결과 군집에 있는 개수를 의미하고, b는 단어 쌍이 다른 정답 군집에 속하고 같은 결과 군집에 있는 개수를 의미한다. 그리고 c는 같은 정답 군집에 있고 다른 결과 군집에 있는 개수를 의미하고, d는 다른 정답 군집에 있고 다른 결과 군집에 있는 개수를 의미한다[23]. 이렇게 <표 5>를 바탕으로 외부 평가 (식 1)-(식 4)을 계산한다[23, 25].

〈표 2〉 이원 분할표

| | 정답 군집이 같을 때 | 정답 군집이 다를 때 |
|---|---|---|
| 결과 군집이 같을 때 | a | b |
| 결과 군집이 다를 때 | c | d |

$$F\text{-}measure = \frac{2 \cdot \frac{a}{a+b} \cdot \frac{a}{a+c}}{\left(\frac{a}{a+b} + \frac{a}{a+c}\right)} \quad \text{(식 1)}$$

$$Rand\ Statistic = \frac{(a+d)}{(a+b+c+d)} \quad \text{(식 2)}$$

$$Jaccard\ Coefficient = \frac{a}{(a+b+c)} \quad \text{(식 3)}$$

$$Folkes\ and\ Mallows = \sqrt{\frac{a}{(a+b)} \cdot \frac{a}{(a+c)}} \quad \text{(식 4)}$$

(식 1)-(식 4)와 같이 이원 분할표를 사용하는 방법은 높은 값이 좋은 성능을 나타내며, 0과 1사이의 값으로 나타난다.

#### 2.4.2. 상대 평가

이 논문에서는 상대 평가 방법으로 *Dunn Index*와 *Davies-Bouldin Index*를 이용할 것이다. 각각의 결과 군집이 $S_i$, $i=1,2,\cdots,n'$일 때, *Dunn Index*과 *Davies-Bouldin Index*는 각각 (식 5)과 (식 6)과 같다.

$$Dunn\ Index = \min_{i=1,\ldots,n'}\left\{\min_{j=i+1,\ldots,n'}\left\{\frac{d(S_i,S_j)}{\max_{k=1,\ldots,n'} diam(S_k)}\right\}\right\} \quad \text{(식 5)}$$

여기서, $diam(S_k)$는 $k$번째 결과 군집 $S$의 지름이고 $d(S_i,S_j)$는 $i$번째 결과 군집과 $j$번째 결과 군집의 최소 거리이다.

$$Davies\text{-}Bouldin\ Index = \frac{1}{n'}\sum_{i=1}^{n'}\max_{i\neq j}\left\{\frac{s_i + s_j}{d_{ij}}\right\} \quad \text{(식 6)}$$

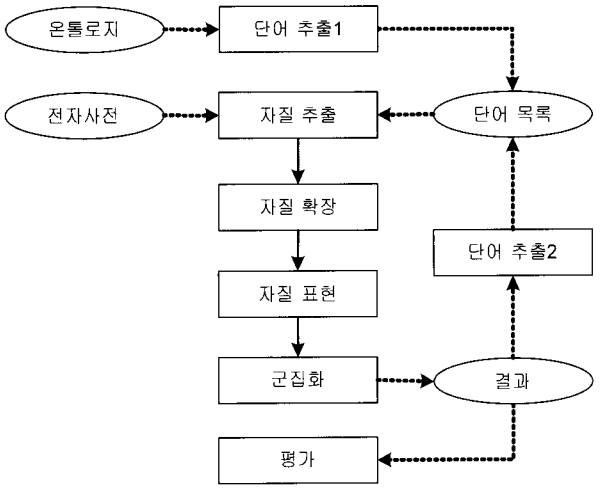
여기서, $d_{ij}$는 $i$번째 군집과 $j$번째 군집의 거리이고 $s_i$는 $i$번째 군집의 중심 값과 $i$번째 군집의 모든 단어와의 평균 거리이다. *Dunn Index*의 방법은 낮은 값이 좋은 성능을 나타내고 *Davies-Bouldin Index*의 방법은 높은 값이 좋은 성능을 나타낸다.

## 3. 단어 군집화 시스템

(그림 1)은 이 논문에서 제안된 단어 군집화 시스템이다. 이하의 절에서는 이 시스템의 각 구성 요소에 대해서 자세히 설명할 것이다.

### 3.1 단어 추출1

이 논문에서 외부 평가를 위해서는 미리 정의된 정답 군집이 있어야 한다. 이를 위해서 이 논문에서는 온톨로지 UWIN을 이용하여 정답 군집을 구축하고 정답 군집에 속한 일부의 단어를 실험 대상 단어로 추출한다. 이 논문은




(그림 1) 단어 군집화 시스템

UWIN 상에서 '배(ship)', '풀', '나무', '꽃', '포유류', '건물'의 6개 단어를 선택하여 그들 단어에 속한 하위 단어 210개를 단어 군집화 대상 단어로 추출한다.

### 3.2 자질 추출

3.1에서 추출된 단어를 군집화하기 위해서는 먼저 각 단어의 자질을 추출하여야 한다. 이 논문은 2.1절에서 언급했듯이 사전의 뜻 풀이말을 이용하여 자질을 추출하며 각 단어(표제어)의 뜻 풀이말로부터 명사와 동사를 추출하여 자질로 이용한다. 예를 들어 군집 대상 단어 '강아지'의 경우 전자사전의 뜻 풀이말은 <표 3>과 같다.

〈표 3〉 '강아지'의 뜻 풀이말

| | |
|---|---|
| 뜻 풀이말 | 강아지 (명) 1) 개의 새끼.<br>강아지 (명) 2) 어린아이를 귀여워해 이르는 말. |
| 뜻 풀이말의 의미 분별 결과 | 강아지 (명) 개_3 +의 새끼_2 +.<br>강아지 (명) 어린아이+를 귀여워하+아 이르_2 +는 말_1 +. |

의미 분별 결과에서 '개_3'은 '개'의 세 번째 의미를 표현하며, 결과적으로 '강아지'의 자질은 '개_3', '새끼_2', '어린아이', '귀여워하', '이르_2', '말_1'이다.

### 3.3 자질 확장

이 논문에서는 자질을 확장하기 위해서 자질 집합에 속한 단어를 다른 단어로 **치환**하는 방법과 기존의 자질에 새로운 단어를 **추가**하는 방법을 사용하며, UWIN 상에서 **상위 단어** (parent word)로 확장하는 방법과 최상위 개념에서 **고정 높이**에 해당하는 단어로 확장하는 방법으로 구분한다.

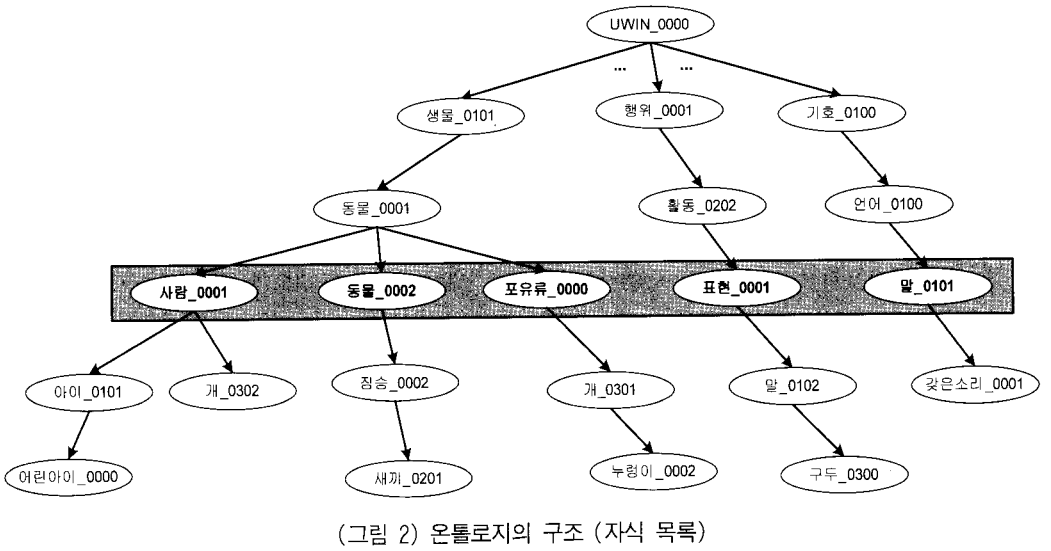
(그림 2)와 같은 온톨로지 구조 상에서 '강아지'의 자질을 상위 단어 추가 방법으로 확장하면 <표 4>와 같다. (그림 2)에서 각 노드의 이름에서 단어 뒤에 사용된 번호는 다의어의 의미번호이다. 예를 들어 '말_0102'는 다의어 의미번호 01번에 세부 의미번호 02를 나타낸다.

(그림 2)와 같은 온톨로지 구조 상에서 고정높이를 3으로 정하여 '강아지'의 자질을 고정높이 단어 추가 방법으로 확장하면 <표 5>와 같다. 고정 높이 3에 해당하는 단어는 최상위 개념인 'UWIN_0000'에서 높이가 3인 단어를 의미하며, (그림 2)에서는 '사람_0001', '동물_0002', '포유류_0000', '표현_0001', '말_0101'이 해당된다.

<표 4>와 <표 5>를 보면, 치환 방법은 원래의 자질에 비해 오히려 자질이 작아지는 것을 알 수 있고, 추가 방법은 자질이 커지는 것을 알 수 있다. 이러한 이유는 뜻 풀이말에서 명사와 동사를 추출하고 치환할 때는 명사만을 고려하기 때문에 동사가 제거되기 때문이다.

### 3.4 자질 표현

자질 표현은 추출된 군집화 대상 단어의 자질을 벡터로 표현한다. 자질을 표현하는 대표적인 방법으로 *tf*idf*, 상호정보량(*mutual information*) 등이 있다[26, 27]. 이 논문에서는 뜻 풀이말 자질이 정보검색에서 검색 대상인 문헌보다 크기가 상대적으로 작기 때문에 *idf* 가중치 기법이 아닌 *df* 기법




(그림 2) 온톨로지의 구조 (자식 목록)

〈표 4〉 '강아지'의 자질 (상위 단어)

| 뜻 풀이말 | 개_3 | 말_1 | 새끼_2 | 어린아이 | 귀여워하, 이르_2 |
|---|---|---|---|---|---|
| 치환 | 앞잡이, 사람, 포유류 | 언어_1, 표현 | 짐승 | 아이_1 | - |
| 추가 | 개_3, 앞잡이, 사람, 포유류 | 말_1, 언어_1, 표현 | 새끼_2, 짐승 | 어린아이, 아이_1 | 귀여워하, 이르_2 |

〈표 5〉 '강아지'의 자질 (고정 높이)

| 뜻 풀이말 | 개_3 | 새끼_2 | 어린아이 | 말_1 | 귀여워하, 이르_2 |
|---|---|---|---|---|---|
| 치환 | 사람, 포유류 | 동물 | 사람 | 말_1, 표현 | - |
| 추가 | 개_3, 사람, 포유류 | 새끼_2, 동물 | 어린아이, 사람 | 말_1, 말_1, 표현 | 귀여워하, 귀여워하 |

을 사용한다[28]. $m$개의 자질 벡터를 $\overrightarrow{w_i}=\langle x_{i1}, x_{i2},..., x_{im}\rangle$ 라고 할 때, *tf*df*는 (식 7)와 같다.

$$x_{ij} = tf_{ij} \times df_j \quad \text{(식 7)}$$

여기서, $tf_{ij}$는 $i$번째 단어의 $j$번째 자질의 개수이고, $df_j$는 $j$번째 자질을 포함하는 단어의 개수이다.

계산 시간을 줄이기 위하여, 자질을 표현할 때 군집 대상 단어 사이에 연관성이 없는 자질(*df*가 1인 자질)을 제거한다. 예를 들어 <표 3>에서 '어린아이', '아이_1', '귀여워하'는 전체 단어 중에서 '강아지'라는 단어에만 존재하는 자질로 이를 제거한다.

### 3.5 군집화

군집화는 방법에 따라 계층적 군집화(hierarchical clustering) 평면적 군집화(partitional clustering), 복합적 군집화(hybrid clustering)로 분류된다[27]. 군집화 알고리즘에 사용되는 유사도 측정 방법은 코사인 계수(*cosine coefficient*), 카이제곱(*chi square*), 자카드 계수(*Jaccard coefficient*) 등이 있다[29]. 이 논문에서는 비교적 성능이 우수한 계층적 군집화 방법 중에 단일 연결(Single-link) 방법[30, 31]으로 단어 군집화를 수행하였고, 유사도 측정 계수로 (식 8)와 같은 코사인 계수를 사용하였다.

$$\cos(\overrightarrow{w_i}, \overrightarrow{w_j}) = \frac{\sum_{z=1}^{m}(x_{iz} \times x_{jz})}{\sqrt{\sum_{z=1}^{m}(x_{iz}^2)\sum_{z=1}^{m}(x_{jz}^2)}} \quad \text{(식 8)}$$

여기서, 분자는 두 단어의 자질 벡터 $i$와 $j$의 내적이고, 분모는 두 단어 벡터 $i$와 $j$의 길이의 곱이다.

### 3.6 단어 추출2

군집화 알고리즘을 수행하면, 자질 확장 방법에 따라 군집의 개수가 다르게 나타날 수 있고(<표 6> 참조), 자질 확장 과정에서 누락되는 단어가 생기기 때문에 군집 대상 단어의 수가 자질 확장 방법에 따라 다르게 나타난다. 객관적인 평가를 위하여, 1차 단어 군집화 후에 나타난 결과에서 군집의 개수가 가장 많고 단어의 개수가 가장 작은 군집을 기준으로 자질 추출, 확장, 표현, 군집화를 수행해서 최종 단어 군집 결과를 낸다.

## 4. 성능 평가

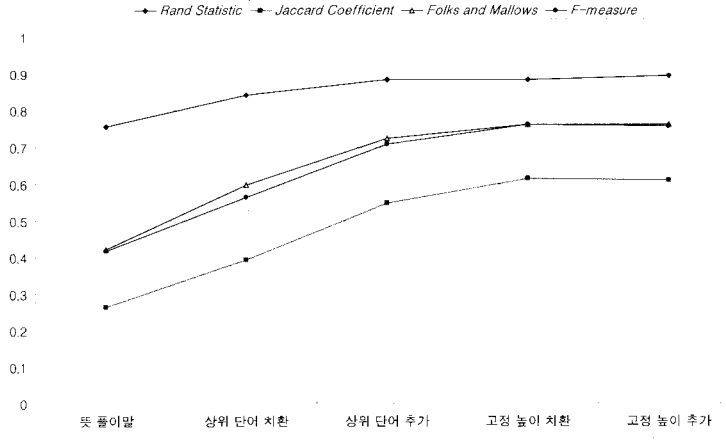
이 장에서는 단어 군집화 시스템을 성능을 평가한다. 실험은 군집화 시스템의 입력 단어의 자질을 각각 뜻 풀이말, 뜻 풀이말의 상위 단어 치환 및 추가, 고정 높이 단어 치환 및 추가 등 모두 5가지 방법으로 자질을 확장하고 각 방법에 대해서 단어 군집화를 수행했을 때의 성능을 비교하고 분석한다.

이 논문에서 온톨로지는 UWIN[16]을 사용하였으며, UWIN에 포함된 단어 목록의 뜻풀이말에 대한 의미분별은 [18]의 결과를 사용하였다. 또한 고정높이는 조정이 가능하나, 이 논문에서는 고정높이를 3으로 고정하여 모든 실험을 수행하였다. 6개의 온톨로지 군집에서 모두 210개의 단어를 추출한 뒤, 군집의 개수를 지정하지 않았을 때, 자질의 확장 방법에 따라 군집의 개수가 <표 6>과 같다.

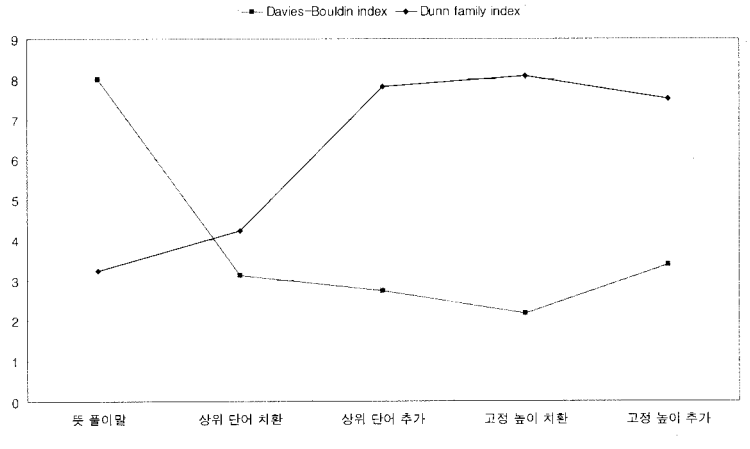
〈표 6〉 군집 개수

| 자질 표현 | 결과 군집 수 |
|---|---|
| 뜻 풀이말 | 19 |
| 상위 단어 치환 | 15 |
| 상위 단어 추가 | 10 |
| 고정 깊이 치환 | 7 |
| 고정 깊이 추가 | 7 |

온톨로지 군집의 개수가 6개인 반면, 뜻 풀이말을 이용했을 때의 결과 군집의 개수가 19개이고 고정 깊이 단어 확장 방법은 정답 군집의 개수에 가까운 7개로 형성되었다. 이러한 이유는 뜻 풀이말 자질의 확장을 통해 공통 자질이 증가하여, 뜻 풀이말만을 사용한 방법에 비해 상대적으로 큰 군집을 형성한다는 것을 알 수 있다. 이 논문에서는 단어 군집 평가의 객관성을 위하여 결과 군집의 개수를 19개로 통일한다. 그리고 최종 결과 군집을 외부 평가와 상대 평가 방법으로 자질 확장 방법에 따른 성능을 비교하고 분석한다. 또한 2.3절에서 언급했듯이 객관적인 평가를 위하여 외부 평가와 상대 평가를 이용한다. 외부 평가 방법으로는 *Rand*




(그림 3) 외부 평가




(그림 4) 상대 평가

*Statistic, Jaccard Coefficient, Folkes and Mallows Index, F-measure*를 사용할 것이고, 상대 평가 방법으로는 *Dunn Index*와 *Davies-Bouldin Index*를 사용할 것이다.

### 4.1 외부 평가

(그림 3)은 외부 평가 방법인 (식 1)-(식 4)으로 평가한 결과이다.

(그림 3)에서 *F-measure*의 값이 뜻 풀이말, 상위 단어 치환, 상위 단어 추가, 고정 높이 치환, 고정 높이 추가 순으로 각각 0.42, 0.56, 0.7, 0.762, 0.758일 때, **상위 단어 치환** 방법은 뜻 풀이말을 사용한 방법보다 25.2%[3]의 성능 향상을 보인다. 이는 단어를 군집화하는데 뜻 풀이말 자체를 사용하는 방법보다 뜻 풀이말을 온톨로지 상의 상위 단어로 치환하는 방법이 좋은 군집을 형성한다고 볼 수 있다. **상위 단어 추가** 방법이 뜻 풀이말을 사용한 방법보다 50%의 성능 향상을 보이고 상위 단어 추가 방법보다 33.1%의 성능 향상을 보인다. 이는 뜻 풀이말을 온톨로지 상의 상위 단어로 확장할 때, 기존의 뜻 풀이말의 명사와 동사를 이용하는 것이 단어 군집화 성능 향상의 요인이라고 판단할 수 있다.

3) ((1-초기값)-(1-실험값))/(1-초기값)*100

**고정 높이 치환** 방법은 뜻 풀이말을 사용한 방법보다 59.2%의 성능 향상을 보이고 상위 단어 추가 방법보다 18.3%의 성능 향상을 보이고 상위 단어 치환 방법보다 45.1%의 성능 향상을 보인다. 이에 반해 **고정 높이 추가** 방법은 뜻 풀이말을 사용한 방법보다 58.4%의 성능 향상을 보이지만 고정 높이 치환 방법에 비해서 1.8%의 성능 손실을 보인다. 온톨로지 상의 고정 높이 단어를 단어 군집화 자질 확장에 사용할 경우, 치환 방법과 추가 방법에는 거의 차이가 없고 오히려 뜻 풀이말을 치환하는 방법이 약간의 성능 저하를 보인다. 이러한 결과로 볼 때, 자질을 확장할 때 사용하는 단어로 뜻 풀이말의 상위 단어보다 고정 높이 단어를 사용하는 방법이 월등히 좋다고 판단된다.

### 4.2 상대 평가

(그림 4)는 각각 상대 평가 방법인 (식 5)과 (식 6)을 이용하여 평가한 결과이다.

**상대 평가**는 단어 군집화의 결과를 절대적인 기준으로 비교할 수는 없지만 상대적으로 뜻 풀이말을 단어 군집화에 사용하는 방법보다 뜻 풀이말을 온톨로지 상의 단어로 확장하는 방법이 좋은 성능을 보인다. 이러한 결과는 앞에서 언급한 외부 평가 결과와 일치한다.

외부 평가와 상대 평가를 종합해 보면, 뜻 풀이말 자체를 단어 군집화의 자질로 사용하는 방법보다 뜻 풀이말을 온톨로지 상의 상위 단어나 고정 높이에 해당하는 단어로 치환 및 추가하는 방법이 좋은 성능을 보인다.

## 5. 결 론

이 논문에서는 사전의 뜻 풀이말이 단어를 함축적으로 가장 잘 표현한다는 사실을 이용하여 사전의 뜻 풀이말을 이용한 단어 군집화 시스템을 설계하고 구현하였다. 그러나 사전의 뜻 풀이말 자체는 매우 함축적으로 단어를 표현하기 때문에 자질이 매우 작은 특징이 있다. 이러한 특징은 뜻 풀이말을 이용한 단어 군집화 결과가 다수의 작은 군집으로 나타난다. 다수의 작은 군집을 양질의 큰 군집으로 만들기 위하여 뜻 풀이말에 추상적인 말이 쓰인다는 특성을 자질 확장에 이용하였다. 여기서 추상적인 말은 온톨로지 상에서 상위 단어에 해당하는 단어로 이 논문에서는 뜻 풀이말의 추상적인 자질을 한 단계 위의 상위 단어로 확장하거나 온톨로지 상에서 어떤 고정 높이에 해당하는 단어로 확장함으로써 단어 군집화 성능을 향상시키는 방법을 제안하였다. 실험 결과, 단어를 군집화할 때 단어의 자질로 뜻 풀이말을 사용한 방법보다 뜻 풀이말에 온톨로지 상의 상위 단어로 추가하는 방법이 50%의 성능 향상을 보였고 고정 높이 단어로 치환하는 방법이 59.2%의 성능 향상을 보였다. 이는 뜻 풀이말을 확장할 때 온톨로지 상의 상위 단어보다 최상위 개념 노드에서 고정 높이에 해당하는 단어를 사용하는 것이 단어 군집화 성능을 크게 향상 시키는 것으로 판단된다. 또한 뜻 풀이말을 온톨로지 상의 상위 단어로 확장할 경우, 동사를 제거하고 명사를 치환하는 방법보다 뜻 풀이말에 상위 단어를 추가하는 방법이 33.1%의 성능 향상을 보였다. 이는 단어를 군집할 때 뜻 풀이말의 동사가 단어의 의미를 구분하는데 도움이 된다고 판단 할 수 있다.

앞으로 이러한 단어 군집화 기법을 사전 검색 시스템이나 정보 검색 시스템과 같은 응용 시스템에 적용하고 다양한 자질 확장을 시도하는 연구로 이어져야 될 것이다.

## 참 고 문 헌


[1] 임영희, "후처리 웹 문서 클러스터링 알고리즘", 한국정보처리학회 논문지 B, Vol.9, No.1, pp.7-16, 2002.

[2] 윤보현, 김현기, 노대식, 강현규, "검색결과의 브라우징을 위한 계층적 클러스터링", 한국정보과학회 논문집, Vol.17, No.1, pp.342-344, 2002.

[3] 최준혁, 전성해, 이정현, "베이지안 SOM과 부트트랩을 이용한 문서 군집화에 의한 문서 순위조정", 한국정보처리학회 논문지, Vol.7, No.7, pp.2108-2115, 2000.

[4] 김건오, 고영중, 서정연, "어휘 클러스터링을 이용한 자동 문서 요약", 한국정보과학회 논문집 B, Vol.29, No.1, pp.464-465, 2002.

[5] Franz, M., McCarley, J. S., Ward, T., and Zhu, W.-J., "Unsupervised and supervised clustering for topic tracking", Proceedings of SIGIR Forum, Vol.24, pp.310-317, 2001.

[6] Shin, S. and Choi, K.-S., "Automatic word sense clustering using collocation for sense adaptation", Proceedings of Global WordNet Conference, pp.320-325, 2004.

[7] 이상훈, 김기태, "클러스터링 기법을 이용한 키워드 유사도 순위화 알고리즘에 따른 사용자 질의 확장", 한국정보과학회 논문집, Vol.30, No.1, pp.479-481, 2003.

[8] Brown, P. F., Della Pietra, V. J., de Souza, P. V., Lai, J. C. and Mercer, R. L. "Class-based n-gram models of natural language", Computational Linguistics, Vol.18, No.4, pp. 467-479, 1992.

[9] Chen, J. N. and Chang, J. S., "Topical clustering of MRD senses based on information retrieval techniques", Computational Linguistics, Vol.24, No.1, pp.61-96, 1998.

[10] The EAGLES Lexicon Interest Group, Preliminary Recommendations on Lexical Semantic Encoding, Final Report EAGLES LE3-4244, 1999.

[11] Federici, S., Montemagni, S., and Pirrelli, V. "Inferring semantic similarity from distributional evidence : An Analogy-based approach to word sense disambiguation", Proceedings of the ACL/EACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. pp.90-97, 1997.

[12] Smadja, F. "Retrieving collocations from text : Xtract", Computational Linguistics, Vol.19, No.1, pp.143-177, 1993.

[13] Lesk, M. "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone", Proceedings of SIGDOC '86, pp.24-26, 1986.

[14] Banerjee, S. and Pedersen, T. "An adapted Lesk algorithm for word sense disambiguation using WordNet", Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Vol.2276, pp. 136-145, 2002.

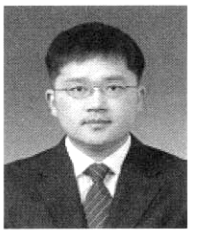
[15] 김준수, 옥철영, "정제된 의미정보와 시소러스를 이용한 동형이의어 분별 시스템", 한국정보처리학회 논문지 B, Vol.12, No.7, pp.829-840, 2005.

[16] 최호섭, 옥철영, "한국어 의미망 구축과 활용 : 명사를 중심으로", 한국어학회, Vol.17, pp.301-329, 2002.

[17] 옥철영, "우리말 개념망 명사 데이터 구축", ETRI 최종연구보고서, 1998.

[18] 조평옥, 안미정, 옥철영, 이수동, "사전 뜻 풀이말에서 구축한 한국어 명사 의미 계층구조", 한국인지과학회 논문지, Vol.10, No.3, pp.1-10, 1999.

[19] 한국과학기술원 전문용어언어공학센터, CoreNet 다국어 어휘망 : 제2권 한국어 어휘 의미망, KAIST PRESS, 2005.
[20] 최석두, 조혜민, "다국어 시소러스의 설계", 한국정보관리학회 학술대회 논문집, Vol.8, pp.5-10, 2001.
[21] 황순희, 윤애선, "워드넷 기반 한국어 명사 어휘의미망의 정제", 한국인지과학회 춘계학술대회 발표논문집, pp.267-272, 2005.
[22] Fellbaum, C., WordNet : An Electronic Lexical Database, MIT Press, 1998.
[23] Halkidi, M. B., and Vazirgiannis, Y. M, "Cluster validity methods : Part I", ACM SIGMOD Record, Vol.31, No.2, pp. 40-25, 2002.
[24] 김정하, 이재윤, "문헌 클러스터링 결과의 성능 평가 방법에 관한 비교 연구", 한국정보관리학회 논문집, Vol.7, pp.45-50, 2000.
[25] Halkidi, M. B. and Vazirgiannis, Y. M, "Cluster validity checking methods : Part II", ACM SIGMOD Record, Vol.31, No.3, pp.19-27, 2002.
[26] Salton, G. and McGill, M. J., Introduction to Modern Information Retrieval, McGraw Hill, 1983.
[27] Patrick, P. Clustering by Committee. Ph.D. Dissertation, Department of Computing Science, University of Alberta, 2003.
[28] 최재혁, 서혜성, 노상욱, 최경희, 정기현, "온톨로지 기반의 웹 페이지 분류 시스템", 한국정보처리학회 논문지 B, Vol.11, No.6, pp.723-734, 2004.
[29] 한승희, 이재윤, "문헌 클러스터링을 위한 유사계수간의 연관성 측정", 한국정보관리학회 논문집, Vol.6, pp.25-28, 1999.
[30] Jain, A. K. and Dubes, R. C., Algorithms for Clustering Data, Prentice-Hall, Inc., 1988.
[31] Johnson, S.C, "Hierarchical clustering schemes", Psychometrika, Vol.2, pp.241-254, 1967.



**박 은 진**

e-mail : bakeunjin@bada.hhu.ac.kr
2003년 한국해양대학교 자동화정보공학부 (학사)
2002년~2004년 (주) 블루코드테크놀로지, 사원
2004년~현재 한국해양대학교 컴퓨터공학과 석사과정
관심분야 : 자연언어처리, 한국어정보처리, 정보검색, 정보추출

**김 재 훈**

e-mail : jhoon@mail.hhu.ac.kr
1986년 계명대학교 전자계산학과(학사)
1988년 한국과학기술원 전산학과(공학석사)
1996년 한국과학기술원 전산학과(공학박사)
1988년~1997년 한국전자통신연구원 선임연구원
2000년~2002년 한국과학기술원 첨단정보기술연구센터 연구원
2001년~2002년 USC, Information Sciences Institute, 방문연구원
1997년~현재 한국해양대학교 컴퓨터공학과 부교수
관심분야 : 자연언어처리, 한국어정보처리, 정보검색, 정보추출

**옥 철 영**

e-mail : okcy@ulsan.ac.kr
1982년 서울대학교 컴퓨터공학과(학사)
1984년 서울대학교 컴퓨터공학과(석사)
1993년 서울대학교 컴퓨터공학과(박사)
1994년 러시아 TOMSK 공과대학 교환교수
1996년 영국 GLASGOW 대학교 객원교수
1984년~현재 울산대학교 컴퓨터정보통신공학부 교수
관심분야 : 한국어정보처리, 의미분별, 온톨로지, 지식베이스, 기계학습, 문서분류