工學碩士 學位論文

한국어 의미 계층망을 이용한 자질 확장에 따른 단어 군집화의 성능 향상

Performance Improvement of Word Clustering
by Extending Features Using Korean Ontology

指導敎授 金 載 熏

2006年 8月

韓國海洋大學校 大學院
컴 퓨 터 工 學 科
朴 殷 珍

本 論文을 朴殷珍의 工學碩士 學位論文으로 認准함

委員長 工學博士 李 章 世 印

委　員 工學博士 柳 吉 洙 印

委　員 工學博士 金 載 熏 印

2006年 6月

韓國海洋大學校 大學院
컴퓨터工學科 朴 殷 珍

목 차

Performance Improvement of Word Clustering
by Extending Features Using Korean Ontology

Eun-Jin Park

Department of Computer Engineering,
Graduate School, Korea Maritime University.

Advised by Jae-Hoon Kim

Abstract

In this thesis, we describe design and implementation of a word clustering system using a definition of an entry word in a dictionary, called a dictionary definition. Generally word clustering needs various features like words and performance of a system for the word clustering depends on using some kinds of features. A 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 or 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 thesis, 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.

제 1 장 서 론

최근 몇 년간 초고속인터넷이 활발히 보급되었다. 우리나라는 지난해까지 경제개발협력기구(OECD) 회원국 중, 초고속인터넷보급률에서 4년 연속 세계 1위를 차지했다. 이런 급속한 인터넷의 보급으로 인하여 인터넷은 많이 대중화되었고 온라인 상에서 획득할 수 있는 정보의 양 또한 기하급수적으로 증가하였다. 인터넷에서 쉽게 접할 수 있는 정보로는 네이버[1]와 같은 각종 포털 사이트의 실시간 인터넷 뉴스[2]에서부터 개인 블로그[3], 다음 까페의 게시판[4], 네이버 지식[5], 싸이월드[6]의 게시판 등에 이르기까지 다양하고 방대하다. 이제는 방대한 정보의 더미 속에서 내가 필요한 정보를 찾아내는 것이 대단히 어려운 문제가 되었다. 비록 네이버, 구글[7] 등과 같은 포털 사이트에서 인터넷 정보검색 엔진을 제공하고, 사용자는 이들의 검색엔진을 통해 찾고자 하는 정보를 쉽게 찾을 수 있게 되었지만, 계속적으로 늘어나는 정보로 인하여 검색된 문서의 수가 지속적으로 증가하고 있어서 검색된 대량의 문서를 활용하는 문제가 발생하였다. 대량의 검색된 문서를 활용하기 위해서는 검색된 결과를 적절히 분류해야 하지만 분류 작업은 천문학적인 비용이 들고 지속적으로 늘어나는 특성 때문에 사람이 하기에 적합하지 않다. 이러한 이유로 최근 많은 정보에서 기계가 스스로 학습하여 연관된 정보를 한 곳에 모으는 기법인 군집화(clustering)에 관한 연구가 활발히 이루어지고

[1] http://www.naver.com/
[2] http://news.naver.com/
[3] http://section.blog.naver.com/
[4] http://cafe.daum.net/
[5] http://kin.naver.com/
[6] http://cyworld.nate.com/
[7] http://www.google.co.kr/

있다[1-3].

군집화 기법은 군집 대상에 따라서 문서 군집화(document clustering)와 단어 군집화(word clustering)로 분류된다. 문서 군집화는 군집 대상이 문서이고 서로 연관이 있는 문서끼리 분류하는 기법을 말하며, 검색결과의 후처리[1-3], 문서 자동 요약[4], 사건 탐색 및 추적[5] 등에 사용되고 있다. 단어 군집화는 군집 대상이 단어이고 유사한 의미를 가진 단어끼리 분류하는 기법을 말하며, 용어의 모호성 해소[6], 정보 검색 시스템의 질의 확장[7] 등에 사용되고 있다.

일반적으로 단어 군집화 시스템은 단어에 대한 자질로 대량의 말뭉치에서 추출한 언어 자원(바이그램(bigram) 정보 혹은 격 정보)을 이용한 방법[8]과 사전의 뜻 풀이말을 이용한 방법[9]이 있다. 대량의 말뭉치에서 추출한 언어 정보를 이용한 방법은 말뭉치에 따라서 단어의 사용빈도가 다르기 때문에 단어 군집화의 결과가 일정하지 않은 문제가 있고, 사용빈도가 낮은 단어에 대해서는 양질의 군집을 형성하기 어렵다는 문제가 있다. 사전의 뜻 풀이말을 이용한 방법은 뜻 풀이말의 표제어 의미를 쉽게 풀어서 설명해 놓은 특성을 이용한다. 예를 들어 사전에서 '강아지'라는 표제어의 뜻 풀이말은 "개의 새끼를 이르는 말"이다. '개'는 표제어 '강아지'의 의미를 포함하는 말이기 때문에 다른 표제어의 뜻 풀이말 중에 '개' 혹은 '새끼'라는 단어가 들어 있는 표제어는 '강아지'와 비슷한 의미로 볼 수 있다. 그러나 사전의 뜻 풀이말에서 '개'나 '새끼'가 들어있는 표제어는 찾기 어렵기 때문에 뜻 풀이말을 이용한 단어 군집화 시스템은 군집화 결과가 다수의 작은 군집으로 나타나고 연관된 단어라도 같은 군집을 형성하지 못하는 문제가 발생한다. 이러한 뜻 풀이말의 자질 부족 문제를 해결하려는 연구가 있었다[10]. 이 연구에서는 대량의

말뭉치로부터 추출한 언어 정보를 이용하여 자질[8] 부족 현상을 해결하려고 하였으나 자질 확장에 사용되는 말뭉치에 따라서 군집화 결과가 다르게 나타나고 사용 빈도가 낮은 단어에 대해서는 자질 확장이 어려운 단점이 있다.

이 논문에서는 사전의 뜻 풀이말을 이용한 단어 군집화 시스템의 자질 부족 문제를 해결하기 위하여 온톨로지(ontology)를 이용한 자질 확장을 통해 이러한 문제를 해결한다. 뜻 풀이말에 나타난 단어(자질)를 온톨로지 상에서 한 단계 위의 단어인 **상위 단어**(parent concept or word)로 확장하거나 최상위 개념에서 특정한 높이에 있는 단어인 **고정 높이 단어**(fixed-depth concept of word)로 확장한다. 그리고 효율적인 자질 확장을 위하여 기존의 뜻 풀이말 자질에 확장 대상 단어를 **추가**하는 방법과 기존의 뜻 풀이말 자질을 확장 대상 단어로 **치환**하는 방법으로 구분하여 자질을 확장한다. 이러한 과정을 통해 부족한 자질 문제를 해결할 수 있고 온톨로지를 이용함으로써 말뭉치를 이용한 방법보다 단어 군집화에 적합한 자질을 얻을 수 있다. 예를 들어 '강아지'의 뜻 풀이말 자질인 '개'는 온톨로지 상에서 '강아지'보다 상위 개념이고 다시 '개'는 '동물'의 하위 개념이다. 즉, 뜻 풀이말에서는 '개'라는 자질이 '강아지'의 의미를 나타내지만, 뜻 풀이말을 온톨로지로 확장하면 '강아지'는 '동물'이라는 의미까지 포함하게 된다. 이렇게 '강아지'와 같이 표제어의 의미가 확장되면 표제어 사이의 공통자질이 생겨나기 때문에 자질 부족 문제를 해결할 수 있다. 이 논문에서는 객관적인 평가를 위하여 사람이 개입한 **외부 평가**(external validation)와 알고리즘 자체의 성능을 평가하는 **상대 평가**(relative validation)로 단어 군집 결과를 평가하였다.

[8] 단어 군집화 알고리즘에서 단어의 의미를 나타내는 요소로서, 일반적으로 뜻 풀이말의 자질은 뜻 풀이말에 나타난 명사 혹은 동사이다.

이 논문은 다음과 같이 구성된다. 2장에서는 이 논문과 관련된 연구를 살펴보고 3장에서는 한국어 의미 계층망을 이용한 자질 확장에 관하여 자세히 설명한다. 4장에서는 자질 확장에 따른 단어 군집화 시스템의 성능을 비교하고 분석한다. 마지막으로 5장에서는 실험 결과를 바탕으로 결론을 맺고 향후 연구 과제를 언급한다.

제 2 장 관 련 연 구

이 장에서는 단어 군집화 시스템에 관한 연구와 군집화 알고리즘, 사전의 뜻 풀이말을 단어 군집화 시스템에 이용하는 연구, 온톨로지에 관한 연구, 군집화 평가 방법에 관한 연구를 간략히 소개한다.

2.1 단어 군집화

기계학습(machine learning)은 시스템 스스로 경험을 쌓아 가면서 관찰하게 되는 다양한 현상 혹은 주어진 대량의 데이터로부터 유용한 지식을 자동으로 추출하는 것을 말한다[11]. 이러한 기계학습은 감독학습(supervised learning)과 자율학습(unsupervised learning)으로 나눌 수 있다. 감독학습은 입력과 정답을 반복적으로 제공함으로써 기계가 스스로 학습하도록 하는 방법을 말하며, 문서분류(document classification)가 여기에 속한다. 자율학습은 입력만 반복적으로 제공함으로써 기계가 스스로 학습하도록 하는 방법을 말하며, 군집화가 여기에 속한다. 따라서 단어 군집화는 입력으로 주어지는 대량의 데이터로부터 의미가 유사한 단어를 하나로 모으는 자율학습 방법이다.

단어 군집화 시스템에서 단어 군집의 기준이 되는 자질을 추출하는 방법은 단어의 용례를 이용한 방법[8]과 사전의 뜻 풀이말을 이용한 방법[9]이 있다. 단어의 용례를 이용하는 방법은 “유사한 의미의 단어는 비슷한 용례를 가진다”라는 가정에 근거한다. 이 방법은 대량의 말뭉치에서 단어의 바이그램 정보 혹은 격 정보를 추출하여 단어 군집화의 군집 대상 단어에 대한 자질로 사용한다. 단어의 바이그램 정보란 어떠한 기준(보통 명사)의 단어 쌍을 말한다. 예를 들어 ‘한국 해양 대학교의 축제’에서 명사인 ‘한국’, ‘해양’,

‘대학교’, ‘축제’의 바이그램 정보는 (‘한국’, ‘해양’), (‘해양’, ‘대학교’), (‘대학교’, ‘축제’)가 된다. 단어의 격 정보는 동사를 기준으로 주격 혹은 목적격의 명사-동사 쌍을 말한다. 예를 들어 “밥을 먹다”에서 동사 ‘먹다’의 목적격 정보는 (‘밥’, ‘먹다’)가 된다. 이 방법은 상대적으로 사용 빈도가 낮은 단어는 말뭉치에서 거의 나타나지 않기 때문에 연관성이 있는 단어라도 같은 군집을 형성하기 어려운 단점이 있다. 단어의 뜻 풀이말을 자질로 이용하는 방법은 “유사한 단어는 비슷한 뜻 풀이말을 가진다”라는 가정에 근거한다[9]. 이 방법에서는 사전의 뜻 풀이말과 온톨로지에서 표제어의 위치 정보를 이용하여 단어 군집을 형성하고, 이를 바탕으로 단어의 모호성을 해소하고 정보 검색에 적용하였다[9].

2.2 군집화 알고리즘

군집화는 방법에 따라 계층적 군집화(hierarchical clustering), 평면적 군집화(partitional clustering), 복합적 군집화(hybrid clustering)로 그림 2.1과 같이 분류된다[12].

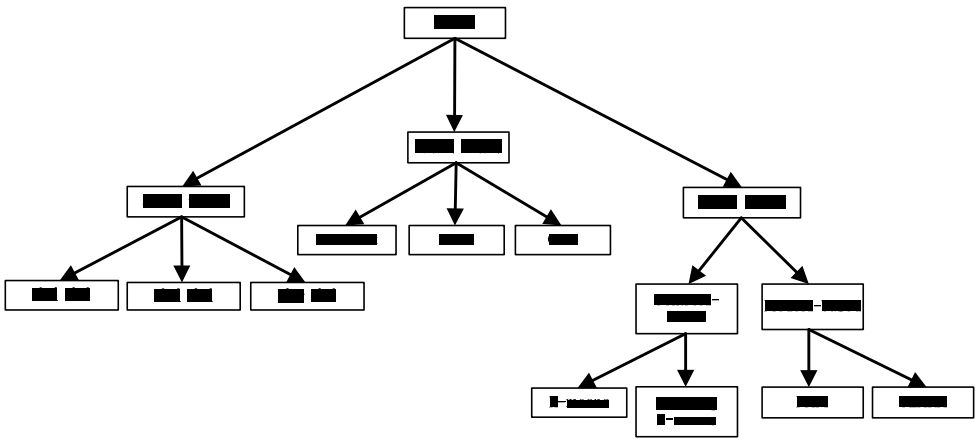


그림 2.1 군집화 알고리즘의 분류

Figure 2.1 Classification of clustering algorithms

다음은 군집화 알고리즘의 일반적인 접근방법이다[13].

1) 병합(agglomerative)과 분할(divisive): 병합이란 각각의 단어를 개별적인 군집으로 보고 가장 가까운 군집을 하나의 군집으로 묶음으로써 새로운 군집을 형성해 나가는 방법이다. 어떤 종료 조건(일반적으로 결과 군집의 개수)이나 하나의 군집이 형성될 때까지 반복적으로 새로운 군집을 형성해 나가는 방법으로 가장 많이 쓰인다. 분할이란 전체의 단어를 하나의 군집으로 보고 서로 다른 성격을 가진 덩어리를 따로 분리해 나가면서 군집을 형성한다. 마찬가지로 어떤 종료 조건이나 각각의 데이터로 분리될 때가지 반복해 나간다.

2) 하드(hard)와 소프트(soft): 하드 군집화 알고리즘은 하나의 단어가 반드시 하나의 군집에 속하는 알고리즘이고, 소프트 군집화 알고리즘은 하나의 단어가 여러 개의 군집에 속하는 알고리즘이다.

1) 계층적 군집화

계층적 군집화는 단어 간의 유사도를 바탕으로 유사도가 높은 것부터 하나씩 계층적으로 군집을 형성해 나가며, 어떤 종료 조건이 만족될 때까지 반복적으로 군집을 형성해 나가는 방법으로 표 2.1과 같다.

표 2.1 계층적 군집화 알고리즘

Table 2.1 The hierarchical clustering algorithm

1) 전체 단어 n개를 n개의 군집으로 초기화한다. 군집 대상의 단어의 수가 n개 존재할 때, 각각의 단어는 하나의 군집에 할당된다. 주어진 단어-자질 행렬($n \times m$ 행렬)에서 단어 간의 유사도를 계산하여 유사도 행렬($n \times n$ 행렬)을 만든다.
2) 유사도 행렬을 바탕으로 가장 유사한 두 개의 군집을 찾아내어 하나의 군집으로 병합한다.
3) 새로 형성된 군집과 기존의 군집들 사이의 유사도를 계산한다.
4) 어떤 종료 조건 혹은 최종적으로 하나의 군집이 형성할 때까지 2번과 3번 과정을 반복한다.

표 2.1에서 유사도 행렬은 단어-자질 행렬을 바탕으로 모든 단어 사이의 유사한 정도를 나타내며 표 2.2와 같다.

표 2.2 유사도 행렬

Table 2.2 A similarity matrix

	w_1	w_2	w_3	$\cdots$	w_n
w_1	0				
w_2	s_{21}	0			
w_3	s_{31}	s_{32}	0		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
w_n	s_{n1}	s_{n2}	$\cdots$	$s_{(n)(n-1)}$	0

표 2.2에서 w_i는 i번째 단어를 의미하고 s_{ij}는 i번째 단어와 j번째 단어 사이의 유사도를 의미한다. 그림 2.2는 계층적 군집화 알고리즘의 군집 형성 단계를 나타내는 계통수(dendrogram)이다.

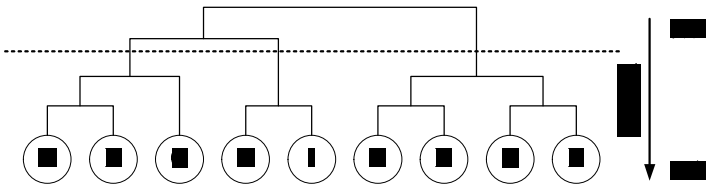
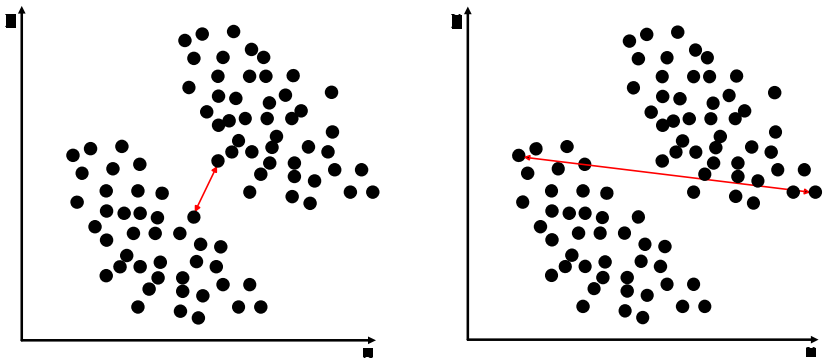


그림 2.2 계층적 군집화 알고리즘의 계통수

Figure 2.2 A dendrogram of hierarchical clustering algorithms

계층적인 군집화에서는 계통수에서 유사도를 기준으로 군집을 추출할 수 있다. 예를 들어 그림 2.2에서 점선을 기준으로 4개의 군집({A, E, C}, {H, I}, {D, B}, {G, F})을 얻을 수 있다. 표 2.1의 계층적 군집화 알고리즘의 3번째 단계인 새로운 군집과 기존의 군집들 간의 유사도를 계산하는 방법에 따라 단일 연결(single-link), 완전 연결(complete-link), 평균 연결(average-link) 알고리즘으로 나눌 수 있다. 단일 연결 알고리즘은 그림 2.3의 (a)와 같이 두 군집 사이의 유사도를 계산할 때 각 군집에 속한 원소들 사이에서 거리가 가장 작은 값을 두 군집 사이의 유사도로 사용하고, 완전 연결 알고리즘은 그림 2.3의 (b)와 같이 두 군집 사이의 유사도를 계산할 때 각 군집에 속한 원소들 사이에서 거리가 가장 큰 값을 사용한다. 여기서 거리는 유사도와 상대적인 개념으로 거리가 가까우면 유사도가 커지고 거리가 멀어지면 유사도가 작아진다.



(a) 단일 연결 (b) 완전 연결

그림 2.3 두 군집 사이의 유사도

Figure 2.3 Similarity between clusters

평균 연결 알고리즘은 그림 2.4와 같이 두 군집 사이의 유사도를 계산할 때 각 군집에 속한 원소들 사이의 평균 거리를 두 군집 사이의 유사도로 사용한다.

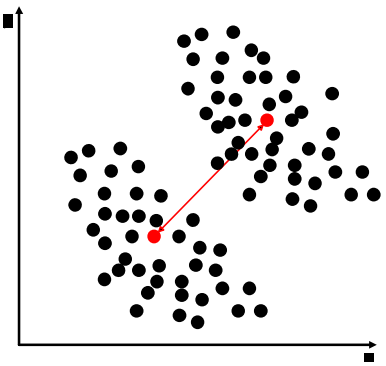


그림 2.4 두 군집 사이의 유사도 (평균 연결)

Figure 2.4 Similarity between clusters (average-link)

2) 평면적 군집화

평면적 군집화는 계층 정보를 생성하지 않는다. 단지 목표 군집의 수 k가 주어지면 분할된 k개의 군집을 형성한다. 일반적으로 평면적 군집화는 초기 시작 위치를 바꿔서 여러 번 실행하여 좋은 결과를 얻어낸다. 평면적 군집화 방법은 계층적 군집화 방법보다 양질의 군집 결과를 얻을 수 없지만 시간 복잡도가 계층적 군집화 방법보다 좋다. 표 2.3은 대표적인 평면적 군집화 알고리즘인 k-means 알고리즘이고, 그림 2.5는 2개 군집의 중심값이 변하는 과정을 나타낸다.

표 2.3 k-means 알고리즘

Table 2.3 The k-means algorithm

1) 임의의 k개의 지점을 정해서 초기 군집의 중심값으로 정한다.
2) 모든 단어에 대하여 중심값에서 가장 가까운 거리에 있는 단어를 하나의 새로운 군집으로 묶는다.
3) 모든 단어가 각 군집에 할당되면 각 군집의 중심값을 다시 계산한다.
4) 2,3 단계를 T회 반복하거나 군집의 중심값이 변하지 않을 때까지 반복한다.

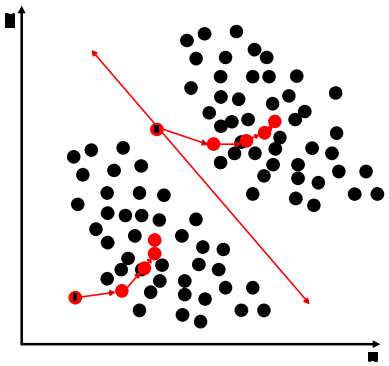


그림 2.5 중심값 변화

Figure 2.5 Movement of centroid

k-means 알고리즘은 최초 중심값의 시작위치에 따라 다른 결과가 나타나므로 초기 중심값의 위치를 정할 때 가능한 고르게 분포할 수 있도록 해야 한다.

계층적 군집화 방법은 계층적인 결과를 출력하고 좋은 성능의 군집을 형성하고 실행 시간이 늦은 단점이 있지만 평면적 군집화 방법은 빠른 실행 시간이 특징이다. 복합적 군집화는 계층적 군집화의 성능과 평면적 군집화의 실용성을 적절히 조합한 방법이다.

2.3 사전의 뜻 풀이말을 단어 군집화에 이용하는 연구

사전 뜻 풀이말의 유사한 정도를 이용하여 단어의 모호성을 제거하는 연구가 있었다[9,14]. 이 연구에서는 뜻 풀이말에 같은 말이 나타나면 서로 연관이 있는 단어로 가정한다. 그러나 사전의 뜻 풀이말 자질의 크기가 작아서 각 단어 사이에 공통 자질이 나타나지 않으면 의미가 유사한 단어라도 같은 군집을 형성하기 어렵다. 이러한 사전의 뜻 풀이말의 자질 부족 문제를 해결하려는 연구가 있었다[10]. 이 연구에서는 사전의 뜻 풀이말을 대량의 말뭉치에서 추출한 언어정보(바이그램 정보 혹은 격 정보)로 확장하여 뜻 풀이말 자질의 크기를 확장하였지만 확장에 사용된 말뭉치에 따라 사용되는 단어의 사용 빈도가 다르기 때문에 확장에 사용되는 말뭉치에 따라 군집화 성능이 다르고 상대적으로 출현빈도가 낮은 단어에 대해서는 자질 확장이 어렵다.

2.4 온톨로지

온톨로지는 실세계에 존재하는 개념들이 서로 어떻게 관련되어 있는가에 대한 지식으로 나무구조(tree)로 표현된다. 온톨로지는 정보검색, 의료정보와 바이오정보, 인공지능 및 에이젠트, 전자상거래, 지능형 인터넷 등 다양한 기술분야에 적용되며, 이미 분야별로 이에 대한 연구가 활발히 진행되고 있다[9]. 현재 국내에서 개발 중인 온톨로지로는 카이스트의 CoreNet[15], 오롬정보의 NexusBase[16], 부산대학교의 KorLex[17], 전자통신연구원의 ETRI 어휘개념망[18], 울산대학교의 UWIN[18,19] 등이 있다. CoreNet의 경우, 개념 기반의 다국어 어휘의미망으로 한국어, 중국어, 일본어로 구축되어 있고, 단일어 사전과 기존의 워드넷을 이용하여 반자동으로 구축되어 있으며, 자연언어처리 및 의미기반 지식처리 시스템에 활용하고 있다. NexusBase는 국제 표준에 맞추어 구축 중인 국내 최대 규모의 온톨로지로써 40만 용어 이상을 포함하고 있고 오롬 시소러스 시스템과 연동되어 있으며 다국어 시소러스 형태로 구축 중이다. KorLex는 한국형 워드넷(WordNet[20])으로서 워드넷의 한국어 번역 결과이다. ETRI 어휘개념망은 개체명 개념망으로 백과사전기반 질의응답 시스템에 활용되고 있다. UWIN은 단어의 뜻 풀이말을 바탕으로 단어의 세부의미 수준까지 계층 분류가 되어 있고, 단어의 모호성 해소 및 형태소 분석 등에 응용되고 있다. 이 논문에서는 단어 군집화 시스템의 성능 평가와 자질 확장에 울산대학교의 UWIN을 사용하였다. 그림 2.6은 온톨로지의 일부를 나타낸 것이다.

[9] http://www.etnews.co.kr/news/detail.html?id=200306230148

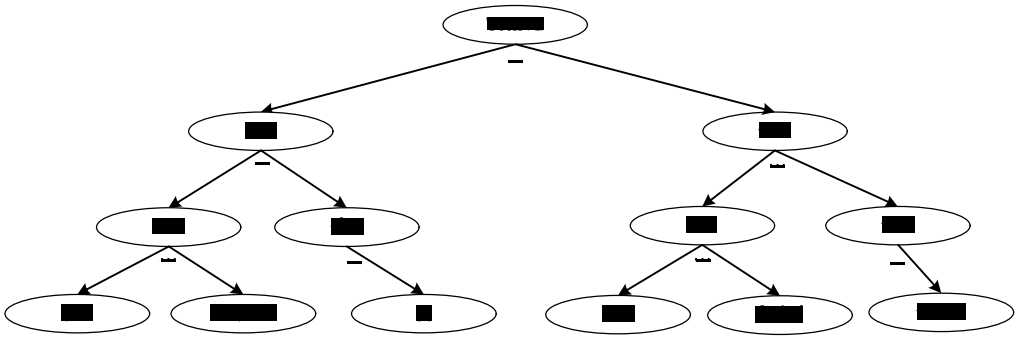


그림 2.6 온톨로지 일부

Figure 2.6 A part of ontology

2.5 군집화 평가 방법

군집화 평가는 접근 방법에 따라 크게 세 가지로 분류된다[21,22]. 사람이 미리 정의한 군집과 기계가 형성한 군집을 비교하는 방법인 **외부 평가**(external validation)와 기계가 형성한 군집 결과를 평가할 때, 외부 기준을 사용하지 않고 데이터 자체만을 이용하여 평가하는 방법인 **내부 평가**(internal validation), 그리고 통계적인 방법을 이용하지 않고 동일한 군집화 알고리즘에서 서로 다른 군집 환경(자질 표현 및 유사도 측정)을 이용하는 방법인 **상대 평가**(relative validation)가 있다. 각각의 대표적인 평가 방법은 표 2.4와 같다[21,23].

표 2.4 군집화 평가 방법

Table 2.4 Methods for clustering validation

외부 평가	*Rand Statistic, Jaccard Coefficient, Folkes and Mallows Index, Huberts Γ Statistic, F-measure*
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 Index, Davies-Bouldin Index*)를 이용하여 단어 군집화 시스템의 성능을 평가한다.

1) 외부 평가

외부 평가 방법은 사람이 미리 정해둔 정답 군집을 이용하기 때문에 사람의 주관에 따른 성능 평가가 가능하다는 특징이 있다. 외부 평가는 표 2.5와 같은 이원 분할표(2×2 contingency table)를 이용하며, 이 논문에서는 *Rand Statistic, Jaccard Coefficient, Folkes and Mallows Index, F-measure*을 이용하여 단어 군집화 시스템의 성능을 평가한다.

표 2.5 이원 분할표

Table 2.5 A 2×2 contingency table

	정답 군집이 같을 때	정답 군집이 다를 때
결과 군집이 같을 때	*a*	*b*
결과 군집이 다를 때	*c*	*d*

표 2.5에서 정답 군집은 사람이 미리 정의한 단어 군집 결과를 의미하고 결과 군집은 시스템이 출력하는 단어 군집 결과를 의미한다. 이원 분할표를 계산하기 위해서 모든 입력 단어의 쌍 (w_i, w_j)에 대해서 그림 2.7과 같은 방법으로 구한다.

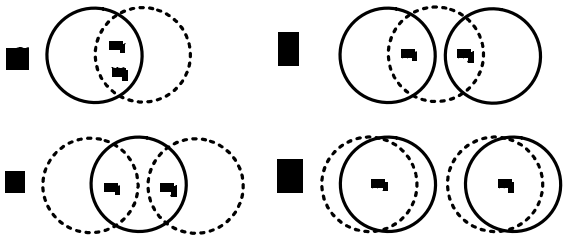


그림 2.7 이원 분할표의 개념도

Figure 2.7 Concept of the 2×2 contingency table

그림 2.7에서 실선 원은 사람이 미리 정의한 정답 군집을 의미하고, 점선 원은 기계가 계산한 결과 군집을 의미한다. a 는 임의의 단어 쌍이 정답 군집 안에 같이 있고 결과 군집 안에 같이 있는 개수이고, b 는 단어 쌍이 다른 정답 군집 안에 있고 같은 결과 군집 안에 있는 개수이다. 그리고 c 는 같은 정답 군집 안에 있고 다른 결과 군집 안에 있는 개수이고, d 는 단어 쌍이 다른 정답 군집 안에 있고 다른 결과 군집 안에 있는 개수이다[21]. 이원 분할표를 바탕으로 외부 평가식 식 (2-1)에서 식 (2-4)를 계산하여 군집화 시스템의 성능을 평가한다[21,24].

$$F\text{-}measure = \frac{2 \cdot precision \cdot recall}{(precision + recall)} = \frac{2 \cdot a}{2 \cdot a + b + c} \tag{2-1}$$

$$Rand\ Statistic = \frac{(a+d)}{(a+b+c+d)} \tag{2-2}$$

$$Jaccard\ Coefficient = \frac{a}{(a+b+c)} \tag{2-3}$$

$$Folkes\ and\ Mallows = \sqrt{\frac{a}{(a+b)} \cdot \frac{a}{(a+c)}} \tag{2-4}$$

식 (2-1)에서 식 (2-4)는 그 값이 높을수록 좋은 성능을 의미하고, 0과 1사이의 값이다.

2) 상대 평가

외부 평가와는 달리 상대 평가는 사람의 개입 없이 군집화 알고리즘에 환경변수(parameter)를 사용하여 출력된 결과 군집 간의 차이를 비교하는 방법이다. 이러한 특성으로 인하여 이 논문에서 제안한 자질 확장에 따른 군집화 성능을 평가할 수 있다. 상대 평가 방법으로 *Dunn Index*와 *Davies-Bouldin Index*를 이용할 것이고, 결과 군집의 개수가 n'일 때, *Dunn Index*와 *Davies-Bouldin Index*는 각각 식 (2-5)와 식 (2-8)과 같다.

$$Dunn\ Index = \min_{i=1,\ldots,n'} \left\{ \min_{j=i+1,\ldots,n'} \left\{ \frac{d_{ij}}{\max_{k=1,\ldots,n'} diam(S_k)} \right\} \right\} \quad (2\text{-}5)$$

여기서, $diam(S_k)$는 k번째 결과 군집 S_k의 지름으로 식 (2-6)과 같고, d_{ij}는 i번째 결과 군집과 j번째 결과 군집의 최단 거리로 식 (2-7)과 같다.

$$diam(S_k) = \max_{w_i, w_j \in S_k} d(w_i, w_j) \quad (2\text{-}6)$$

$$d_{ij} = \min_{w_i \in S_i, w_j \in S_j} d(w_i, w_j) \quad (2\text{-}7)$$

$$Davies\text{-}Bouldin\ Index = \frac{1}{n'} \sum_{i=1} \max_{i \neq j} \left\{ \frac{v_i + v_j}{d_{ij}} \right\} \quad (2\text{-}8)$$

여기서, d_{ij}는 i번째 군집과 j번째 군집의 최단 거리이고 v_i는 i번째 군집의 중심값과 그 군집에 속한 모든 단어와의 평균 거리이다. d_{ij}를 그림으로 나타내면 그림 2.8과 같고 v_i는 식 (2-9)와 같고 그림으로 나타내면 그림 2.9와 같다.

$$v_i = \frac{1}{|S_i| \cdot |S_i - 1|} \sum_{w_i, w_j \in S_i} d(w_i, w_j) \quad (2\text{-}9)$$

여기서 S_i는 i번째 결과 군집이고 $|S_i|$는 S_i 군집의 단어 수이다.

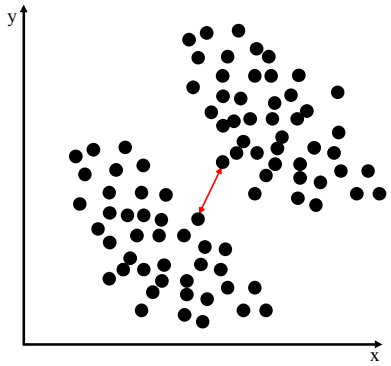


그림 2.8 두 군집 사이의 최단 거리

Figure 2.8 Minimum distance between two clusters

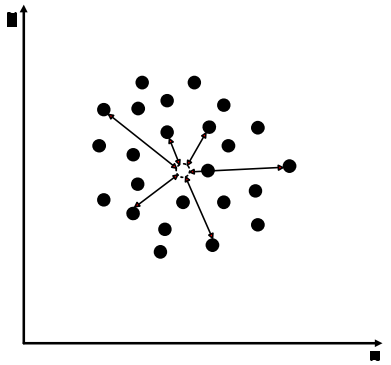


그림 2.9 중심값과 평균 거리

Figure 2.9 Average distance among members

*Dunn Index*는 값이 높을수록 성능이 좋고 *Davies-Bouldin Index*는 값이 낮을수록 성능이 좋다[23].

제 3 장 한국어 의미 계층망을 이용한 자질 확장

이 장에서는 한국어 의미 계층망을 이용한 자질 확장 방법을 구체적으로 기술하고자 한다. 자질 확장은 군집화 시스템의 한 모듈이며, 군집화 시스템 내에서 중요한 역할을 담당한다. 실험에 사용한 단어 군집화 시스템의 구조는 그림 3.1과 같으며 다음과 같은 과정으로 처리된다.

1) 온톨로지에서 단어를 추출하여 군집화 대상 단어 목록을 만든다.
2) 전자사전에서 군집 대상 목록에 해당하는 표제어의 뜻 풀이말을 추출한다.
3) 추출된 자질을 온톨로지를 이용하여 확장한다.
4) 확장된 자질을 벡터로 표현한 뒤, 각 표제어 간의 유사도를 계산한다.
5) 계산된 유사도를 바탕으로 군집화 알고리즘을 수행하여 결과 군집을 형성한다.
6) 객관적인 평가를 위하여 각 확장 방법에 따라 다른 군집 결과를 통일한다. 이를 위해 군집 결과에서 표제어를 다시 추출하고 가장 많은 군집의 수를 알아낸 뒤에 다시 단계 2부터 단계 5까지 수행한다.
7) 외부 평가와 상대 평가 방법으로 최종 결과 군집을 평가한다.

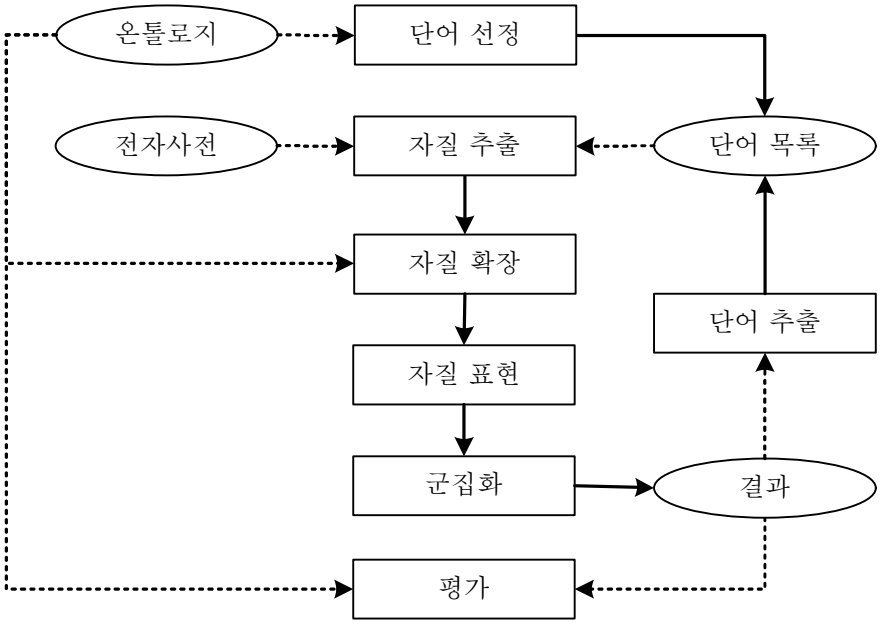


그림 3.1 제안된 단어 군집화 시스템

Figure 3.1 The architecture of our proposed word clustering system

그림 3.1에서 사각형은 시스템 모듈을 의미하고, 타원은 모듈에 사용된 자원을 의미하고 점선은 각 모듈의 입력을 나타내며, 실선은 모듈의 출력을 나타낸다. 이하의 절에서는 이 시스템의 각 구성 요소에 대해 자세히 설명할 것이다.

3.1 단어 선정

단어를 군집화하기 위해서는 군집 대상 단어를 선정해야 한다. 실험에 사용한 군집화 대상 단어는 온톨로지 상에서 '배(ship)', '포유류', '건물', '풀', '나무', '꽃' 의 하위 단어를 사용하였다. '배', '포유류', '건물' 군집은 서로 연관이 적은 군집이고 '풀', '나무', '꽃' 군집은 서로 연관이 큰 군집이다. 따라서 전자는 후자에 비해 좋은 군집화 성능을 보일 것으로 예상된다. 군집화 대상 단어로 모두 304개가 선정되었고 선정된 단어는 표 3.1과 같다. 표

3.1에서 '풀_02'은 표준국어대사전[10]을 기준으로 '풀'의 여러 가지 의미에서 두 번째 의미를 나타내며, 이 논문에서는 자질로 의미가 분별된 단어를 사용한다[11].

표 3.1 군집화 대상 단어

Table 3.1 A word list for clustering

풀_02	갈대 감국 감자_01 **개박하** 건초_01 고구마 고란초 고사리 **고추_01** 구절초 국화_05 **궁궁 나물_01** 나팔꽃 **난_06** 난초_03 달맞이꽃 담배 당근 **도라지_01** **독초_01** 들깨 땅콩 마_02 마늘 **맨드라미 면화_03 모_01** 목화_02 무_02 **미나리_01** 민들레 **밀_03** 바나나 백합_03 **봉숭아** 분꽃 사탕수수 **산채_01** 삼_03 상추_01 샐비어 생강 **선인장_02 수박_01** 수선화 **수세미_01** 수수_01 수초_03 시금치 쑥_02 **쓴풀 야채** 약초 **약풀** 양배추 양파 여러해살이풀 여물_01 연_015 **오이_01** 오이풀 옥수수 용담_04 인삼 자리공 잔디 **잡초_02** 참깨 **참외_01** 채송화 코스모스_01 콩나물 토마토 파초_02 팥 패랭이 풍란 할미꽃 **해바라기_02** 호박_01
배_02	**감시선_01 건착선** 곤돌라 공모선 구축함 **군함_02** 나룻배 나무배 뗏목 만선_03 보트 상선_06 어선_05 여객선 **연락선_01** 운반선 유람선 유조선 **잠수함_02** **트롤선 함정_02** 화물선
건물_03	가옥_01 **가정집_01** 강당 객줏집 거미집 겹집 **고궁_01** 고치_01 곡창_01 공관_02 관가_01 **관사_04 광_01** 교당 국고 국회의사당 궁_05 궁궐 궁전_05 기념관 너와집 노인정 **농원** 누각_02 **단층_01 닭장 대궐_02** 대웅전 마천루 막_05 **막사_02 모기장** 문고_01 미술관 민가 벌집 법당 별장_03 **별채_02** 병동 **병영_01** 본관_04 본당_01 본산 불사_06 브리핑 빌딩 빌라_02 사당_06 **사랑채** 사옥_03 사찰_02 사택_03 산장_03 **새장 생가_01** 성당_03 시골집 **신사_11** 신전_11 아파트 안채_01 **암자_01** 역사_12 **영화관_01** 오두막 오막살이 온실 왕궁 움막 음악당 의사당 **자가_01** 자택 저택_02 **전시관** 절_01 주막 **주막집** 주택 지하실 집_01 **집채_01** 창고_01 천막 체육관 **초가_03 초가집 초당_02** 카페 **캠프** 텐트 판잣집 하숙집 한옥 **행랑채** 헛간 회관

[10] http://www.korean.go.kr/uw/dispatcher/search/dictionary/dic_sear.html
[11] 울산대학교 의미 부착 뜻 풀이말

표 3.1 군집화 대상 단어 (계속)

Table 3.1 A word list for clustering (continued)

꽃_01	국화_01 **매화_02** 산화_05 **연꽃 연화_10 함박_01**
포유류	강아지 개_03 고슴도치 고양이 곰_03 기린_02 **기마_01 꽃사슴 낙타_02** 늑대 다람쥐 돼지 말_05 말승냥이 멧돼지 박쥐 **백마_01** 범_01 **불곰** 사자_011 **산달_03 산돼지_01** 생쥐 소_03 양_05 얼룩말 얼룩소 여우_01 염소_01 원숭이 족제비 쥐_02 코끼리 토끼 호랑이 **흰쥐**
나무_01	가로수 가시나무 감나무 거목_01 고목_01 **관목_04** 나도밤나무 **낙락장송** 낙엽수 너도밤나무 느티나무 **단풍_01** 대_02 대추나무 **도토리나무** 돌배나무 동백나무 **딸기** 떡갈나무 레몬 마로니에 **머루_01** 모과나무 **모란_01** 묘목_01 미루나무 **밀감** 밤나무 **백송_01** 버드나무 **버들_01** 뽕나무 살구나무 **상록수_02** 상수리나무 **서리_13 석남_01** 소나무 **싸리_01 야자_02 야자수** 오미자 옻나무 은행나무 잡목 잣나무 **장작** 전나무 진달래 **차_09 참나무** 콜라 **포플러** 플라타너스 **해송_01**

3.2 자질 추출

3.1절에서 선정된 단어를 군집화하기 위해서는 각 단어의 자질을 추출하여야 한다. 여기서 자질이란 단어 사이의 특징을 구별할 수 있는 요소를 말한다. 예를 들어 표 3.2의 '강아지' 뜻 풀이말에서 '강아지'의 특징을 나타내는 요소로는 '개_03', '의', '새끼_02', '어린아이', '를', '귀여워하', '아', '이르_02', '는', '말_01'이다. 그러나 '의', '를', '아', '는'과 같은 조사는 '강아지'의 뜻 풀이말뿐만 아니라 다른 단어의 뜻 풀이말에도 흔히 나타나는 요소이기 때문에 '강아지'의 고유한 특징을 나타내지 않는다. 그러므로 특징을 잘 나타내지 않는 단어를 제거하면 '개_03', '새끼_02', '어린아이', '귀여워하', '이르_02', '말_01'이 '강아지'의 자질이 된다. 이러한 과정을 **자질 추출**이라 하고 실험에는 사전의 뜻 풀이말의 명사와 동사만 자질로 사용한다.

표 3.2 '강아지'의 뜻 풀이말

Table 3.2 The definition of 'puppy'

뜻 풀이말	1) 개의 새끼. 2) 어린아이를 귀여워해 이르는 말.
뜻 풀이말의 의미 분별 결과	1) 개_03+ 의 새끼_02+ . 2) 어린아이+ 를 귀여워하+ 아 이르_02+ 는 말_01+ .
뜻 풀이말 자질	개_03 새끼_02 어린아이 귀여워하 이르_02 말_01

표 3.2에서 뜻 풀이말의 의미 분별 결과에서 '+'는 같은 어절 안에 형태소를 구분하는 기호이다.

3.3 자질 확장

각 단어의 자질로 사용되는 뜻 풀이말의 크기가 작기 때문에 유사한 의미를 가진 단어라도 공통 자질이 없어서 같은 군집을 형성할 수 없는 문제가 나타난다. 예를 들어 '동물'이라는 공통 의미를 가진 '강아지'와 '고양이'라도 표 3.2와 표 3.3에서처럼 뜻 풀이말 자질에 공통 자질이 없다. 이러한 자질 부족 현상을 해결하기 위하여, 단어의 자질을 다른 말뭉치를 이용하여 공통 자질을 늘리는 것을 **자질 확장**이라고 한다.

표 3.3 '고양이'의 뜻 풀이말

Table 3.3 The definition of 'cat'

뜻 풀이말	1) 고양이과의 짐승. 2) 송곳니가 발달되어 있고 밤눈이 밝아 쥐를 잘 잡음.
뜻 풀이말의 의미 분별 결과	1) 고양이+ 과_04+ 의 짐승+ . 2) 송곳니+ 가 발달+ 되_05+ 아 있_03+ 고 밤눈_02+ 이 밝_02+ 아 쥐_02+ 를 잘 잡_01+ 음+ .
뜻 풀이말 자질	고양이 과_04 짐승 송곳니 발달 밤눈_02 쥐_02 잡_01

이 논문에서 제안한 뜻 풀이말의 자질 확장 방법을 다양한 각도에서 살펴보기 위하여 확장 대상 단어에 따라 **상위 단어**와 **고정 높이 단어**로 구분하였고, 확장 방법에 따라 **치환**과 **추가**로 구분한다. 상위 단어란 사전의 뜻 풀이말에 나온 명사를 기준으로 온톨로지 상의 상위 노드(parent's node)를 의미하고, 고정 높이 단어는 온톨로지의 최상위 노드에서 어떤 특정한 높이에 있는 단어를 의미한다. 실험에서 자질 확장에 사용한 고정 높이는 3이다. 그림 3.1에서 고정 높이 3에 해당하는 단어는 '사람_0001', '동물_0002', '포유류_0000', '표현_0001', '말_0101'이고 다의어 의미 번호까지만 사용하기 때문에 '사람', '동물', '포유류', '표현', '말_01'이 된다. 예를 들어 '새끼_02'의 상위 단어는 '짐승'이고 고정 높이 단어는 '동물'이다.

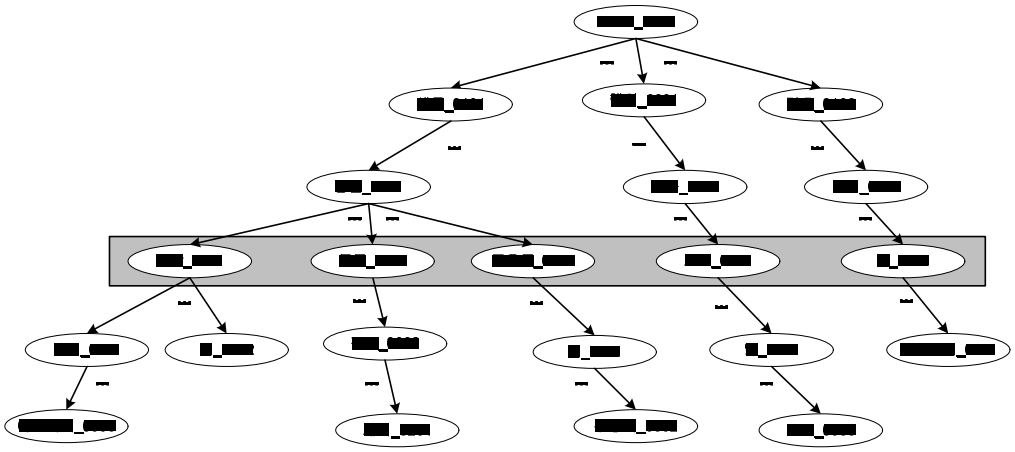


그림 3.1 온톨로지 UWIN의 일부

Figure 3.1 The part of the ontology UWIN

그림 3.1에서 'UWIN_0000'는 온톨로지의 최상위 노드를 나타내고 각 노드에 사용된 번호는 표준국어대사전을 기준으로 작성된 다의어 의미번호이다. 예를 들어 '말_0102'은 표준국어대사전에서 여러 가지 '말'의 의미 중에서 1번째

뜻의 2번째 의미를 나타낸다. 표준국어대사전에서 '말_01'의 뜻 풀이말이 표 3.4와 같을 때, '말_0102'의 의미는 "음성 기호로 생각이나 느낌을 표현하고 전달하는 행위 또는 그런 결과물"이다. 실험에 사용한 사전의 뜻 풀이말에는 다의어 의미번호까지 부착되어 있기 때문에 뜻 풀이말 자질을 확장할 때에는 다의어 의미번호만을 사용한다.

표 3.4 '말_01'의 뜻 풀이말

Table 3.4 The first definition of the 'word'

말_01 [말 :] 「명」
「01」 사람의 생각이나 느낌 따위를 표현하고 전달하는 데 쓰는 음성 기호.
「02」 음성 기호로 생각이나 느낌을 표현하고 전달하는 행위 또는 그런 결과물.
「03」 일정한 주제나 줄거리를 가진 이야기.
「04」 단어, 구, 문장 따위를 통틀어 이르는 말.
「05」 소문이나 풍문 따위를 이르는 말.
「06」 다시 강조하거나 확인하는 뜻을 나타내는 말.
「07」 '망정이지'의 뜻을 나타내는 말.
「08」 '-을 것 같으면'의 뜻을 나타내는 말.
「09」 어떤 행위가 잘 이루어지지 않음을 탄식하는 말.
「10」 앞에서 언급한 사실을 강조하여 말하는 뜻을 나타내는 말.
「11」 어감을 고르게 할 때 쓰는 군말.

이와 같은 방법으로 '강아지'의 뜻 풀이말 자질을 상위 단어로 확장하면 표 3.5와 같고 고정 높이 단어로 확장하면 표 3.6과 같다.

표 3.5 ‘강아지’의 자질 (상위 단어 확장)

Table 3.5 Features of ‘puppy’ under extension of the parent word

뜻 풀이말	개_03	새끼_02	어린아이	말_01	귀여워하, 이르_02
치환	앞잡이, 사람, 포유류	짐승	아이_01	언어_01, 표현	-
추가	개_03, 앞잡이, 사람, 포유류	새끼_02, 짐승	어린아이, 아이_01	말_01, 언어_01, 표현	귀여워하, 이르_02

표 3.6 ‘강아지’의 자질 (고정 높이 단어 확장)

Table 3.6 Features of ‘puppy’ under extension of fixed-depth word

뜻 풀이말	개_03	새끼_02	어린아이	말_01	귀여워하, 이르_02
치환	사람, 포유류	동물	사람	말_01, 표현	-
추가	개_03, 사람, 포유류	새끼_02, 동물	어린아이, 사람	말_01(2), 표현	귀여워하(2), 이르_02

표 3.6에서 괄호 안의 수는 자질의 출현 빈도이고 괄호가 없는 것은 한번 나타난 것이다.

치환 방법은 뜻 풀이말 자질에 비해서 자질의 수가 줄어드는 것을 알 수 있고, 추가 방법은 자질의 수가 늘어나는 것을 알 수 있다. 이것은 사전의 뜻 풀이말에서는 명사와 동사를 추출하지만 확장할 때에는 명사만 확장하기 때문에 확장 방법 중 치환 방법에서는 동사가 사라져서 자질의 수가 줄어든다.

3.4 자질 표현

군집화 대상 단어의 자질을 계산 가능한 벡터 형태로 표현하고 단어-자질 행렬로 나타내는 것을 **자질 표현**이라고 한다. 표 3.7은 *n*개의 군집화 대상 단어와 *m*개의 자질 행렬을 나타낸다.

표 3.7 단어-자질 행렬

Table 3.7 A word-feature matrix

	f_1	f_2	$\cdots$	f_m
w_1	x_{11}	x_{12}	$\cdots$	x_{1m}
w_2	x_{21}	x_{22}	$\cdots$	x_{2m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
w_n	x_{n1}	x_{n2}	$\cdots$	x_{nm}

자질을 표현하는 대표적인 방법으로는 *tf*df*, *tf*idf*, 상호정보량(*mutual information*) 등이 있다[12,25]. 사전의 뜻 풀이말 자질의 수가 문서 자질의 수보다 상대적으로 매우 작기 때문에 *df* 가중치 기법을 사용한다[25]. *m*개의 자질 벡터를 $\vec{w}_i =< x_{i1}, x_{i2}, ..., x_{im} >$라고 할 때, *tf*df*는 식 (3-1)과 같다.

$$x_{ij} = tf_{ij} \times df_j \qquad (3\text{-}1)$$

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

계산 시간을 줄이기 위하여, 자질을 표현할 때 군집 대상 단어 사이에 겹치지 않는 자질(*df*가 1인 자질)을 제거한다. 예를 들어 표 3.5에서 '어린아이', '아이_01', '귀여워하'는 전체 단어의 자질 중에서 '강아지'라는

단어에만 존재하는 자질이기 때문에 이런 자질을 제거한다. 표 3.8은 '강아지'와 '고양이'의 단어-자질 행렬의 일부를 나타낸다.

표 3.8 '강아지'와 '고양이'의 자질 행렬

Table 3.8 The word-feature matrix of 'puppy' and 'cat'

자질 / 단어	고양이	과_04	개_03	짐승	송곳니	새끼_02	…
…	…	…	…	…	…	…	…
강아지	0	0	1	0	0	1	…
고양이	1	1	0	1	1	0	…
…	…	…	…	…	…	…	…

3.5 군집화

실험에 사용한 군집화 알고리즘은 계층적 군집화 방법 중에서 단일 연결 알고리즘이고 군집화 결과는 하나의 단어가 하나의 군집에 속하는 하드 군집을 형성하며 병합 방법으로 군집을 형성한다[26,27]. 단일 연결 알고리즘은 유사도 행렬이 그림 3.2와 같을 때, 유사도가 가장 높은 D와 E 군집을 새로운 군집인 DE로 만든다. 이때 새로운 군집 DE와 다른 군집 A, B, C 사이의 유사도는 기존의 군집 D와 군집 E의 유사도에서 가장 큰 값을 새로운 군집 DE와 다른 군집 A, B, C 사이의 유사도로 사용한다. 예를 들어 그림 3.2에서 군집 A와 군집 D의 유사도는 2이고 군집 A와 군집 E의 유사도는 3이다. 그러므로 새로운 군집 A와 새로운 군집 DE의 유사도는 3이 된다.

	A	B	C	D	**E**
A	0	5	6	2	3
B	5	0	7	4	8
C	6	7	0	1	5
D	2	4	2	0	**9**
E	3	8	5	9	0

ㄹ

	A	B	C	DE
A	0	5	6	3
B	5	0	7	8
C	6	7	0	5
DE	3	8	5	0

그림 3.2 군집 사이의 유사도 계산 과정 (단일 연결)

Figure 3.2 Similarity computation between two clusters (single-link)

일반적으로 군집 사이의 유사도 측정 방법은 코사인 계수(cosine coefficient), 카이제곱(χ^2), 자카드 계수(Jaccard coefficient) 등이 있다[28]. 실험에 사용한 유사도 측정 계수는 식 (3-2)와 같은 코사인 계수이다.

$$\cos(\vec{w}_i, \vec{w}_j) = \frac{\sum_{z=1}^{m}(x_{iz} \times x_{jz})}{\sqrt{\sum_{z=1}^{m} x_{iz}^2} \times \sqrt{\sum_{z=1}^{m} x_{jz}^2}} \qquad (3\text{-}2)$$

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

3.6 단어 추출

군집화 알고리즘을 수행하면, 표 4.1과 같이 자질 확장 방법에 따라 군집의 개수가 다르게 나타날 수 있다. 표 3.1에서 선정된 단어 중에서 '개박하', '쓴풀', '행랑채'와 같은 단어는 사전에 없는 단어이고 이러한 단어는 자질을

표현할 때 사라진다. 온톨로지 상에는 있지만 사전에 나타나지 않은 이유는 온톨로지를 구축할 때 사용한 사전과 이 논문에서 사용한 사전이 다르기 때문이다. 객관적인 평가를 위하여, 1차 단어 군집화 후에 나타난 결과에서 군집의 개수가 가장 많고 단어의 개수가 가장 작은 군집을 기준으로 단어 군집 대상 단어 목록을 다시 작성한다. 따라서 표 3.1에서 굵은 표시의 단어를 제외한 210개의 단어를 기준으로 자질 추출, 확장, 표현, 군집화를 수행해서 최종 결과 군집을 형성한다.

제 4 장 성 능 평 가

이 장에서는 각 자질 확장 방법에 따른 단어 군집화 시스템의 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. 군집화 알고리즘의 목표 군집의 개수를 6개로 지정하고 모두 210개의 단어를 군집화했을 때, 자질 확장 방법에 따른 결과 군집의 개수가 표 4.1과 같이 나타난다.

표 4.1 각 자질 확장 방법의 결과 군집의 수

Table 4.1 The number of word clusters of each method for feature extension

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

목표 군집의 수가 6개인 것은 온톨로지 상에서 6개의 군집(‘배(ship)’, ‘풀’, ‘나무’, ‘꽃’, ‘포유류’, ‘건물’)의 하위 단어를 추출하였기 때문이다. 표 4.1에서 뜻 풀이말을 이용했을 때의 결과 군집의 개수가 가장 많은 19개이고, 고정 높이 단어 확장 방법은 목표 군집의 개수에 가까운 7개로 형성된다. 이것은 뜻 풀이말 자질 확장을 통해 공통 자질이 증가하여, 뜻 풀이말만을 사용한

방법에 비해 상대적으로 큰 군집을 형성했다는 것을 알 수 있다. 표 4.2는 고정 높이 치환 방법으로 자질을 확장하였을 경우의 단어 군집화 결과이다.

표 4.2 단어 군집화 결과(고정 높이 치환)

Table 4.2 Clustering result of the fixed-depth replacement

1	**관목_04 밤나무 잣나무 가시나무 느티나무 전나무 콜라 플라타너스**
2	**구축함 곤돌라 거미집 말_05 미루나무 배_02 보트 뽕나무 양배추 여러해살이풀 유조선 은행나무** 천막 콩나물 **패랭이**
3	**너도밤나무 나도밤나무 감나무 대추나무 돌배나무 떡갈나무 모과나무 동백나무 상수리나무 소나무 옻나무 풍란**
4	**묘목_01 가로수 거목_01 고목_01** 원숭이 나무배 다람쥐 **잡목**
5	불사_06 강당 곡창_01 공관_02 관사_04 주택 국회의사당 궁_05 궁궐 궁전_05 누각_02 막_05 대웅전 집_01 땅콩 마천루 문고_01 민가 벌집 법당 별장_03 가옥_01 건물_03 자택 객줏집 겹집 고치_01 교당 기념관 너와집 사당_06 안채_01 의사당 자리공 저택_02 카페 판잣집 하숙집 한옥 병동 본관_04 본산 생쥐 본당_01 빌딩 빌라_02 사옥_03 사찰_02 사택_03 산장_03 성당_03 시골집 아파트 역사_12 오두막 오막살이 온실 왕궁 움막 음악당 절_01 주막 지하실 창고_01 체육관 텐트 헛간 회관
6	수초_03 사탕수수 감국 담배 목화_02 민들레 백합_03 삼_03 샐비어 생강 토마토 마_02 할미꽃 쑥_02 감자_01 잔디 국화_05 달맞이꽃 마늘 양파 들깨 수선화 바나나 인삼 호박_01 상추_01 나팔꽃 오이풀 연_15 시금치 코스모스_01 갈대 구절초 분꽃 수수_01 용담_04 참깨 팥 고란초 건초_01 고사리 레몬 **버드나무 살구나무** 고구마 당근 **마로니에 무_02** 산화_05 약초 여물_01 옥수수 대_02 진달래 채송화 파초_02
7	얼룩소 강아지 곰_03 개_03 고슴도치 고양이 기린_02 난초_03 늑대 돼지 말승냥이 멧돼지 박쥐 범_01 사자_11 소_03 얼룩말 여우_01 염소_01 족제비 쥐_02 코끼리 토끼 포유류 호랑이
8	여객선 만선_03 운반선 유람선 화물선 나룻배
9	신전_11 양_05
10	브리핑
11	상선_06
12	국고

표 4.2 단어 군집화 결과(고정 높이 치환) (계속)

Table 4.2 Clustering result of the fixed-depth replacement (continued)

13	국화_01
14	낙엽수
15	어선_05
16	노인정
17	뗏목
18	미술관
19	오미자

표 4.2에서 진한 글씨의 단어는 '나무_01'에 속하는 단어이고, 1~4번 군집은 온톨로지에서 추출한 단어 목록인 표 3.1의 '나무_01' 군집과 비슷하고, 5번의 경우에는 '건물_03' 군집과 비슷하고, 6번은 '풀_02' 군집과 비슷하고, 7번은 '포유류' 군집과 비슷하다는 것을 알 수 있다. 10~19번의 단어는 군집화 알고리즘 목표 군집의 수를 19개로 설정하였기 때문에 이런 단어는 다른 단어와 공통 자질이 부족하여 군집을 형성하기 전에 알고리즘 수행이 종료되어서 군집을 형성하지 못한 단어이다.

단어 군집 평가의 객관성을 위하여 결과 군집의 수가 가장 큰 19개를 목표 군집의 개수로 통일하여 군집화 알고리즘을 다시 수행한다. 최종적으로 출력된 결과 군집을 외부 평가와 상대 평가 방법으로 단어 군집화 시스템의 성능을 평가한다. 외부 평가 방법으로는 *Rand Statistic, Jaccard Coefficient, Folkes and Mallows Index, F-measure*를 사용하였고, 상대 평가 방법으로는 *Dunn Index*와 *Davies-Bouldin Index*를 사용하였다.

4.1 외부 평가

그림 4.1은 외부 평가 방법인 식 (2-1)에서 식 (2-4)로 평가한 결과이다.

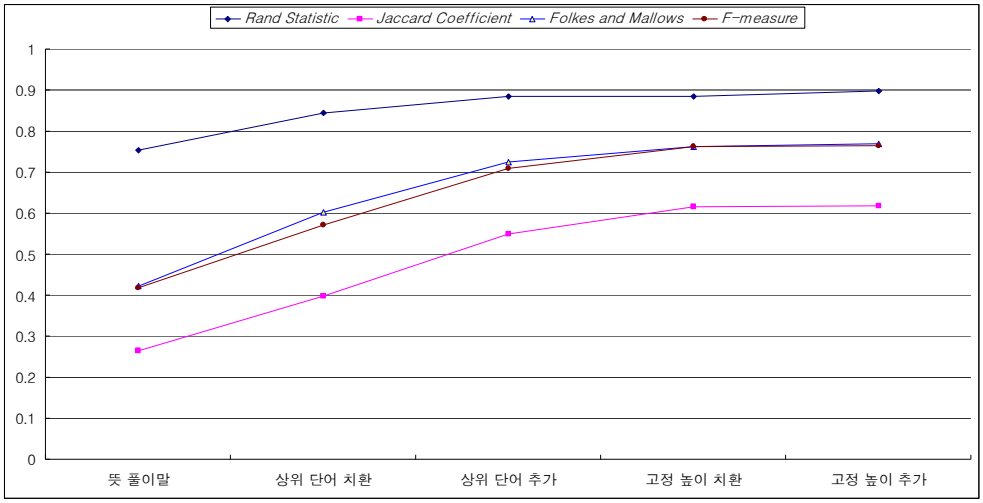


그림 4.1 외부 결과

Figure 4.1 The result of external validation

그림 4.1에서 *F-measure*의 값이 뜻 풀이말, 상위 단어 치환, 상위 단어 추가, 고정 높이 치환, 고정 높이 추가 순으로 각각 0.41, 0.57, 0.70, 0.76, 0.76이고, **상위 단어 치환** 방법은 뜻 풀이말을 사용한 방법보다 26.0%[12]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 49.9%의 성능 향상을 보이고, 상위 단어 치환 방법보다 32.3%의 성능 향상을 보인다. 이는 뜻 풀이말을 온톨로지 상의 상위 단어로 확장할 때, 기존의 뜻

[12] ((1-기준값)-(1-비교값))/(1-기준값)*100
※ 기준값: 뜻 풀이말의 값, 비교값: 상위 단어 치환/추가값 및 고정 높이 치환/추가값

풀이말의 명사와 동사를 이용하는 것이 단어 군집화 성능을 향상시킨다고 볼 수 있다. **고정 높이 치환** 방법은 뜻 풀이말을 사용한 방법보다 59.1%의 성능 향상을 보이고, 상위 단어 추가 방법보다 18.2%의 성능 향상을 보이고, 상위 단어 치환 방법보다 44.6%의 성능 향상을 보인다. 그리고 **고정 높이 추가** 방법은 뜻 풀이말을 사용한 방법보다 59.4%의 성능 향상을 보인다. 온톨로지 상의 고정 높이 단어를 단어 군집화 자질 확장에 사용할 경우, 치환 방법과 추가 방법에는 거의 차이가 없었다. 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 4.2은 각각 상대 평가 방법인 식 (2-5)와 식 (2-8)를 이용하여 평가한 결과이다.

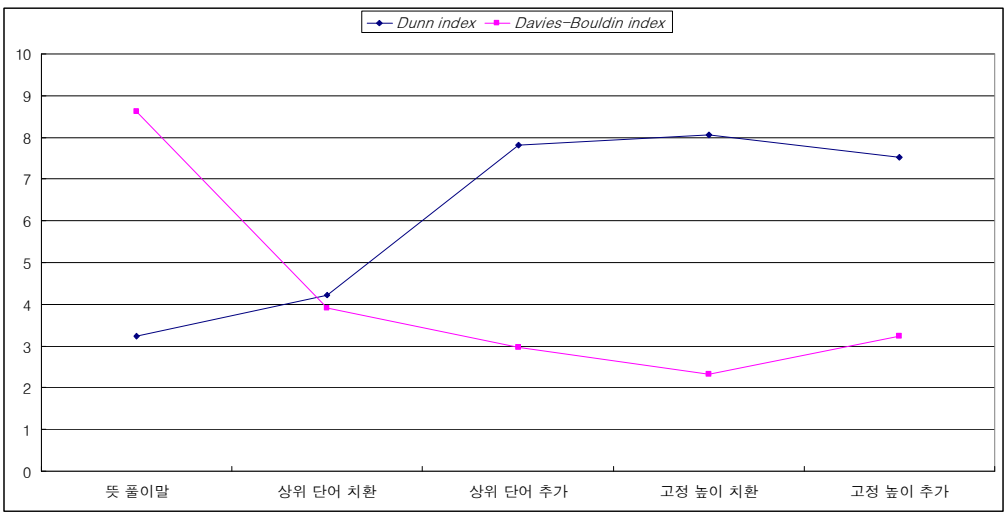


그림 4.2 상대 평가 결과

Figure 4.2 The result of relative validation

상대 평가는 단어 군집화의 결과를 절대적인 기준으로 비교할 수는 없지만 상대적으로 뜻 풀이말을 단어 군집화에 사용하는 방법보다 뜻 풀이말을 온톨로지 상의 단어로 확장하는 방법이 좋은 성능을 보인다는 것을 알 수 있다. 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.

4.3 군집 결과의 타당성

군집이 올바르게 형성되는지 관찰하기 위하여 실험에 사용한 군집 대상 단어를 3개의 집단으로 나누어 군집 대상 단어에 따른 군집화 성능 변화를 알아보았다. 비교 집단 1은 전체 6개 군집(‘배(ship)’, ‘풀’, ‘나무’, ‘꽃’, ‘포유류’, ‘건물’)으로 구성하였고, 집단 2는 단어 간의 특성이 식물로서 유사한 의미를 가진 3개 군집(‘풀’, ‘나무’, ‘꽃’)으로 구성하였다. 마지막 집단 3은 단어 간의 의미가 명확히 구분이 되는 4개 군집(‘배(ship)’, ‘나무’, ‘포유류’, ‘건물’)으로 구성하였다. 그림 4.3은 뜻 풀이말 자질을 고정 높이 단어 추가 방법으로 확장하였고 이를 사용하여 단어 군집 결과를 *F-measure*로 측정한 결과이다.

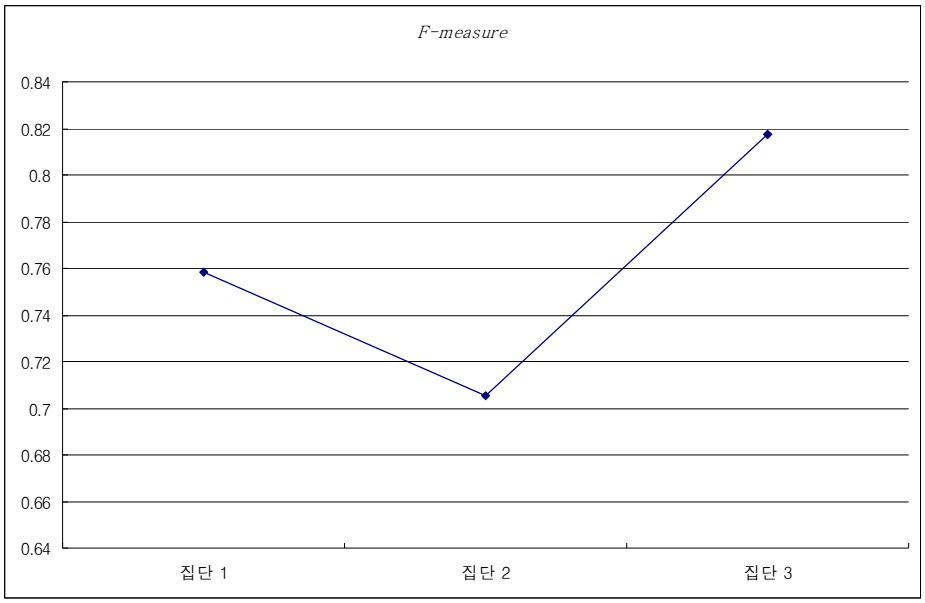


그림 4.3 군집 대상 단어에 따른 성능 변화

Figure 4.3 Performance variation among input word groups

그림 4.3에서 보면 의미가 명확히 구분이 되는 집단 3이 가장 좋은 성능을 보였고, 유사한 의미를 가진 집단 2가 좋지 않는 성능을 보였다. 실험에 사용한 집단 1이 사람이 구분하기 명확한 단어와 명확하지 않은 단어가 적절히 조합된 집단으로 집단 2보다 좋은 결과를 보였다. 이 결과는 유사한 의미를 가진 단어들은 군집화 알고리즘에서도 서로 구별하기 어렵다는 사실을 보여 주고 있다.

제 5 장 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. 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 49.9%의 성능 향상을 보였고 고정 높이 단어로 치환하는 방법이 59.1%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 향상시키는 것을 알 수 있었다. 그리고 뜻 풀이말을 온톨로지 상의 상위 단어로 확장할 경우, 동사를 제거하고 명사를 치환하는 방법보다 뜻 풀이말에 상위 단어를 추가하는 방법이 32.3%의 성능 향상을 보였다. 이는 단어를 군집할 때 뜻 풀이말의 동사가 단어의 의미를 구분하는데 도움이 됨을 알 수 있었다.

이 논문에서는 뜻 풀이말에 의미가 분별 된 사전을 사용하였지만, 앞으로는

의미가 분별 되지 않은 웹 사전을 이용하는 방법과 단순히 뜻 풀이말을 이용하여 군집화 성능을 향상시키는 연구뿐만 아니라 뜻 풀이말에 나오는 예문을 사용하는 방법, 그리고 뜻 풀이말과 예문을 같이 사용하는 방법 등 다양한 방법으로 단어 군집화 시스템의 성능을 향상시키는 연구로 이어져야 될 것이다. 또한 이 논문에서는 자질 치환에 의해 사라지는 단어는 고려하지 않았지만 앞으로는 온톨로지의 크기를 확장하거나 다른 방법으로 자질을 보강하여 단어가 사라지는 문제를 해결하는 등의 연구로 이어져야 할 것이다.

참 고 문 헌

[1] 임영희, "후처리 웹 문서 클러스터링 알고리즘", *한국정보처리학회 논문지*, Vol. 9(B), 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*, Vol. 29(B), 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] The EAGLES Lexicon Interest Group, "5.1 Word Clustering", *Preliminary Recommendations on Lexical Semantic Encoding*, pp. 171-176, 1999.

[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] 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.

[11] Tom M. Mitchell, *Machine Learning*, McGraw Hill, 1997.

[12] Patrick, P. *Clustering by Committee*. Ph.D. Dissertation, Department of Computing Science, University of Alberta, 2003.

[13] Jain, A. K.; Murty, M. N.; and Flynn, P. J., “Data clustering: a review.”, *ACM Computing Surveys*, Vol. 3, pp. 264-323, 1999.

[14] 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.

[15] 한국과학기술원 전문용어언어공학센터, CoreNet 다국어 어휘망: 제2권 한국어 어휘 의미망, KAIST PRESS, 2005.

[16] 최석두, 조혜민, "다국어 시소러스의 설계", *한국정보관리학회 학술대회 논문집*, Vol. 8, pp. 5-10, 2001.

[17] 황순희, 윤애선, "워드넷 기반 한국어 명사 어휘의미망의 정제", *한국인지과학회 춘계학술대회 발표논문집*, pp. 267-272, 2005.

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

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

[20] Fellbaum, C., *WordNet: An Electronic Lexical Database*, MIT Press, 1998.

[21] Halkidi, M. B., and Vazirgiannis, Y. M, "Cluster validity methods: Part I", *ACM SIGMOD Record*, Vol. 31, No. 2, pp. 40-25, 2002.

[22] 김정하, 이재윤, "문헌 클러스터링 결과의 성능 평가 방법에 관한 비교 연구", *한국정보관리학회 논문집*, Vol. 7, pp. 45-50, 2000.

[23] Halkidi, M. B. and Vazirgiannis, Y. M, "Cluster validity checking methods: Part II", *ACM SIGMOD Record*, Vol. 31, No. 3, pp. 19-27, 2002.

[24] Salton, G. and McGill, M. J., *Introduction to Modern Information Retrieval*, McGraw Hill, 1983.

[25] 최재혁, 서혜성, 노상욱, 최경희, 정기현, "온톨로지 기반의 웹 페이지 분류 시스템", *한국정보처리학회 논문지*, Vol. 11(B), No. 6, pp. 723-734, 2004.

[26] Jain, A. K. and Dubes, R. C., *Algorithms for Clustering Data*, Prentice-Hall, Inc., 1988.

[27] Johnson, S.C, "Hierarchical clustering schemes", *Psychometrika,* Vol. 2, pp. 241-254, 1967.

[28] 한승희, 이재윤, "문헌 클러스터링을 위한 유사계수 간의 연관성 측정", *한국정보관리학회 논문집*, Vol. 6, pp. 25-28, 1999.