용어 클러스터링의 성능 평가

박은진[1] 김재훈[1] 옥철영[2]
[1]한국해양대학교 컴퓨터공학과 자연언어처리연구실
[2]울산대학교 컴퓨터정보통신공학부 언어처리연구실
{bakeunjin, jhoon}@hhu.ac.kr[1], okcy@ulsan.ac.kr[2]

Performance Evaluation for Word Clustering

Eun-Jin Park[1] Jae-Hoon Kim[1] Cheol-Young Ock[2]
[1]Department of Computer Engineering, Korea Maritime University.
[2]School of Computer Engineering & Information Technology, Ulsan University

요 약

이 논문에서는 전자 사전의 뜻 풀이말을 이용하여 용어를 자동 분류하는 용어 클러스터링 시스템을 설계하였다. 클러스터링 성능에 영향을 미치는 요소로 자질 선택, 자질 표현 그리고 유사도 측정 등이 있다. 이 논문에서는 이러한 요소들이 용어 클러스터링에 미치는 영향을 평가해보았다. 클러스터링 결과를 객관적으로 비교하기 위해서 용어 클러스터링 결과와 한국어 의미 계층망에서 추출한 정답 클러스터를 비교하였다. 실험 결과, 용어의 뜻 풀이말만 자질로 사용한 방법보다는 뜻 풀이말 자질을 확장하는 방법이 훨씬 더 좋은 결과를 보였다.

1. 서 론

인터넷에서 접근 가능한 정보가 기하급수적으로 늘어난 만큼 방대한 정보를 분류해서 적절한 정보를 찾아내는 시간도 비례하여 증가하였다. 이런 방대한 정보더미 속에서 사용자가 원하는 자료를 찾아서 걸러주는 일을 기계적으로 처리하는 연구가 1960년대 처음 소개 되었던 기계학습이다[3]. 기계학습(Machine Learning)이란 "시스템 스스로가 경험을 쌓아가면서 관찰하게 되는 다양한 현상들 혹은 주어진 다량의 데이터로부터 유용한 지식을 자동으로 추출하는 것"을 말한다. 이러한 기계학습법은 기계가 학습을 하는데 있어서 사람의 개입여부에 따라 지도 학습법과 자율 학습법으로 나눌 수 있다.

클러스터링은 자율 학습 방법이다. 클러스터링은 기계가 방대한 정보를 사람의 개입 없이 자동으로 분류하는 기법이다. 이러한 클러스터링 기법은 다시 클러스터링 대상이 무엇이냐에 따라서 문서 클러스터링과 용어 클러스터링 등으로 나눌 수 있다. 문서 클러스터링은 연관이 있는 문서를 하나의 클러스터로 형성하는 것을 말하며, 용어 클러스터링은 의미가 유사한 용어를 하나의 그룹으로 형성하는 것을 말한다. 용어 클러스터링 시스템은 용어의 모호성 해소, 정보 검색 시스템의 질의어 확장, 문서 요약 시스템 등에 응용되어 사용되었고, 또한 알고리즘의 성능을 높이기 위하여 알고리즘의 전처리 혹은 후처리 단계에서 사용된다[2,6].

이러한 클러스터링의 성능에 영향을 미치는 요소로 자질의 추출 및 선택, 자질의 표현, 그리고 유사도 계산 방법 등이 있고, 이들이 시스템에 미치는 영향을 비교하는 연구가 활발히 진행되고 있다[4,6,8]. 이들 연구에서는 자질과 자질 사이의 연관성을 측정하는데 사용되는 유사계수에 따라 문서 클러스터링 결과를 비교하였으나 용어 클러스터링 시스템은 문서 클러스터링 시스템과 비교했을 때 용어에 대한 자질이 매우 작은 특징이 있어서 용어 클러스터링에 적합한 환경요소를 비교하는 연구가 필요하다.

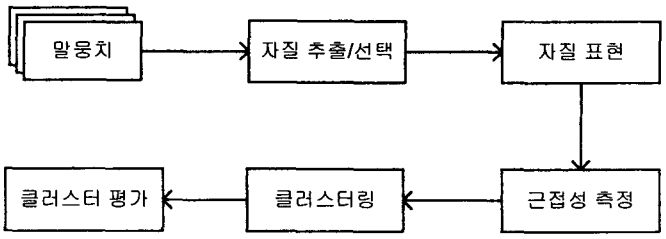
이 논문에서는 자질에 민감한 용어 클러스터링 시스템에 영향을 주는 자질 추출 및 선택 방법, 자질 확장 방법, 자질 표현 방법, 그리고 유사도 계수를 서로 다르게 하여 클러스터링을 수행하고 클러스터링의 결과를 한국어 의미 계층망과 비교하여 어떠한 요소가 클러스터링 성능에 영향을 미치는지 연구하였다.

이 논문은 다음과 같이 구성된다. 2장에서는 클러스터링 시스템에 대해서 설명한다. 그리고 3장에서는 실험 과정 및 결과를 보여주고, 4장에서는 결론 및 향후 과제에 대해 언급한다.

2. 클러스터링 시스템

일반적으로 클러스터링 시스템은 자질 추출 및 선택, 자질 표현, 유사도 측정, 클러스터링 그리고 클러스터의

평가 순으로 이루어지며, [그림 1]과 같다[9].



[그림 1] 시스템 구성

이번 장에서는 이들 구성에 대해 살펴본다.

2.1. 자질 추출 및 선택

말뭉치가 주어지면 클러스터를 형성하는데 유용한 자질을 추출하게 되는데, 일반적으로는 자질을 그대로 사용하지만, 용어 클러스터링의 경우, 각 용어에 대한 자질의 수가 상대적으로 문서 클러스터링(Text Clustering)보다는 작기 때문에 자질을 확장하는 방법이 연구되었다[11].

• 자질 확장 (Gloss Vector)

Gloss Vector 방법은 용어의 자질을 다른 말뭉치를 이용하여 확장시키는 방법이다[11].

예를 들어 '학교'라는 용어의 자질 확장 과정을 살펴보면, '학교'의 뜻 풀이말 <표 1>에서 가능한 자질(주로 명사)를 추출하면 '일정', '목적', '설비', '제도', '규칙', '교사', '피교육자', '교육', '기관'이 된다. 이렇게 추출된 자질은 다른 말뭉치를 통해서 확장된다. <표 2>는 <표 1>에서 추출한 자질 중 '교육'을 포함하는 Bigram의 일부를 나타내며, 이를 자질 '교육'에 대한 확장된 자질로 간주한다. 이런 방법으로 나머지 자질에 대하여 자질을 확장한다. <표 3>은 '학교'에 대해서 자질이 확장된 Gloss Vector의 일부를 보여준다. 이 Gloss Vector가 '학교'의 확장된 새로운 자질 벡터이다.

<표 1> '학교'의 뜻 풀이말

학교 (명) **일정**한 **목적**, **설비**, **제도** 및 **규칙**에 의거해, **교사**가 계속적으로 **피교육자**에게 **교육**을 실시하는 **기관**.

<표 2> 교육을 포함하는 Bigram

교육	학교	36
교육	민족	15
교육	식민지	15
교육	위하	15
교육	나라	14
교육	자녀	12
교육	컴퓨터	11
교육	읽기	11
교육	논리	10
교육	국민	10

<표 3> 학교 뜻 풀이말의 Gloss Vector

공기어 / 뜻풀이말	학교	민족	식민지	위하	나라	자녀
목적	1	9	0	4	2	17
설비	5	13	0	3	5	0
제도	23	14	1	3	2	3
규칙	4	4	3	3	7	5
교사	5	5	4	2	5	5
피교육자	0	6	1	6	8	3
교육	36	15	15	15	14	12
기관	18	10	0	7	8	4
Gloss vector	**92**	**76**	**24**	**43**	**51**	**49**

2.2. 자질 표현

추출된 자질을 계산이 용이한 형태인 벡터로 표현하게되는데 [그림 2]와 같은 행렬 형태로 표현한다.

	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}

[그림 2] 용어-자질 행렬

일반적으로 널리 사용되는 자질 표현 방법으로는 *tf*idf*와 상호정보량(Mutual Information) 등이 있다.

• *tf*idf*

특정 자질이 어떤 용어의 뜻 풀이말에 나타난 빈도수(Term Frequency)를 자질의 벡터로 사용하는 경우도 있지만, 임의의 자질이 전체 용어 뜻 풀이말에 나타난 빈도수를 나타내는 문서 빈도수(Document Frequency)의 역수를 곱한 값을 사용한다[9]. 역문서빈도(Inverse Document Frequency)는 (식 1)과 같다.

$$idf_j = \log \frac{N}{df_j} \qquad \text{(식 1)}$$

여기서, N은 전체 용어의 개수이고, df_i는 i번째 자질이 전체 용어 뜻 풀이말에 나타난 빈도수를 나타낸다. 이때 자질 값은 (식 2)와 같다.

$$x_{ij} = tf_{ij} \times idf_j \qquad \text{(식 2)}$$

여기서, tf_{ij}는 i번째 용어의 j번째 자질의 빈도수이고, idf_j는 j번째 자질의 역문서빈도수(Inverse Document Frequency)이다.

• 상호 정보량 (Mutual Information)

상호 정보량는 정보이론 기반에 의한 연관계수 측정법 중 하나로 용어 w_i에서 자질 f_j에 대한 정보의 양을

측정하는 척도이다. 즉, w_i와 f_j사이의 의존 관계를 정량적으로 나타낸다.

$$x_{ij} = MI(w_i, f_j) = \log \frac{p(w_i, f_j)}{p(w_i)p(f_j)} \quad \text{(식 3)}$$

여기서 f_j는 j번째 자질이고, $p(w_i, f_j)$는 w_i와 f_j의 공기 확률이다. 그리고 $p(w_i)$와 $p(f_j)$는 각각 w_i와 f_j의 독립확률이다.

2.3. 유사도 측정

임의의 두 용어 간의 유사한 정도를 측정하는 방법에는 두 용어가 유사한가를 나타내는 유사도 측정 방법과 얼마나 멀리 떨어져 있는가를 나타내는 거리 측정 방법이 있다. 유사한 정도와 거리는 서로 상대적인 개념이므로 서로 바꿔 사용할 수 있지만 일관성을 위하여 혼합하여 사용할 수는 없다. [그림 3]은 용어-자질 행렬에서 용어 간의 유사도를 측정한 용어간 유사도 행렬이다.

	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_{nn-1}	0

[그림 3] 용어 간 유사도 행렬

이러한 유사도 측정 방법에는 벡터의 내적을 이용한 방법과 코사인 계수를 이용한 방법을 많이 사용한다.

- 내적 (Inner Product)

간단히 두 용어 w_i와 w_j의 유사도를 측정 방법으로 벡터의 내적을 사용한다. 벡터의 내적은 (식 4)과 같다.

$$s_{ij} = Inner(w_i, w_j) = \sum_{z=1}^{m} (x_{iz} \times x_{jz}) \quad \text{(식 4)}$$

여기서 x_{iz}는 i번째 용어의 z번째 있는 자질 값을 나타낸다.

- 코사인 계수 (Cosine Coefficient)

두 용어 w_i와 w_j의 유사도를 측정하는 데 있어서 (식 5)과 같은 코사인 계수를 사용한다.

$$s_{ij} = Cos(w_i, w_j) = \frac{\sum_{z=1}^{m} (x_{iz} \times x_{jz})}{\sqrt{\sum_{z=1}^{m} x_{iz}^2} \times \sqrt{\sum_{z=1}^{m} x_{jz}^2}} \quad \text{(식 5)}$$

여기서 분모는 벡터의 내적이고, 분자는 각각의 자질 벡터의 거리이다.

2.4. 클러스터링

일반적으로 클러스터링 알고리즘은 계층적인 (Hierarchical) 방법, 분할적인 (Partitional) 방법, 그리고 복합적인(Hybrid) 방법으로 분류된다[9]. 계층적인 알고리즘은 용어 간의 유사도을 계산하여 유사도가 높은 것부터 하나씩 계층적으로 클러스터를 재구성하는 방법을 말하며, 분할적인 알고리즘은 유사도가 높은 것을 한데 묶어나가며, 어떤 종료 조건이 될 때까지 분할 혹은 병합함으로써 최적의 클러스터를 하나씩 형성해 나가는 방법이다. 복합적인 알고리즘은 계층적인 알고리즘의 양질성과 분할적인 알고리즘의 실용성을 적절히 조합하는 방법이다. 일반적으로 계층적 알고리즘이 분할적 알고리즘보다는 성능은 우수하지만 속도가 느린 단점이 있다.

2.5. 클러스터 평가

대표적인 자율 학습법(Unsupervised Learning)으로 분류되는 클러스터링의 성능평가는 상대적으로 지도 학습법에 속하는 정보검색이나 범주화 기법(Classification)의 성능평가보다는 어려운 점이 있다. 정보검색이나 범주화 기법에서 성능평가 방법으로 사용되는 정확률(Precision)과 재현율(Recall)은 각 용어에 대한 적합한 질의나 적합한 법주가 미리 정해져 있어서 결과에 대한 객관적이고 절대적인 평가가 가능하다. 그러나 클러스터링의 경우에는 생성된 클러스터가 어느 클러스터(범주)에 해당하는지, 용어가 어느 클러스터로 자동 분류되는지에 대한 판정이 어렵다. 그래서 동일한 환경에서 상대적인 평가를 통해 클러스터링의 성능을 평가한다. 여기서 동일한 환경이란 클러스터링 알고리즘 적용과 관련한 클러스터 수, 자질 추출 방법, 자질 표현 방법, 유사도 측정 방법 등을 말하며, 이를 통일하여 클러스터링 결과를 평가한다. 이 논문에서는 클러스터내의 불순한 정도를 나타내는 엔트로피(Entropy)와 임의의 용어가 가지는 클러스터의 대상후보를 나타내는 퍼플렉스티(Perplexity)를 측정한다[4].

- 엔트로피(Entropy)

정보이론에서 비롯된 엔트로피는 임의의 집합에서 데이터의 동질성을 측정하는 용도로 널리 쓰인다[4,11]. 엔트로피를 계산하는 식은 (식 6)와 같다.

$$H(C_i) = \sum_{x \in C_i} \sum_{s \in S_j} P(x,s) \log_2 P(x,s) \quad \text{(식 6)}$$

여기서 C_i은 기계가 수행한 i번째 클러스터를 의미하고, S_j은 한국어 의미 계층망에서 추출한 j번째 클러스터를 의미한다. $P(x,s)$는 S_j에 C_i가 속할 확률을 나타낸다. 엔트로피가 0에 가까울수록 클러스터링 결과가 우수하다는 의미로 해석된다.

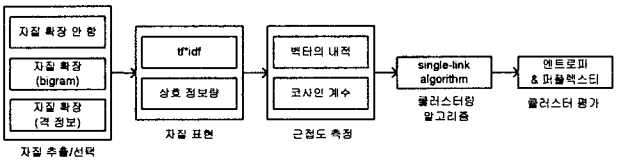
- 퍼플렉스티(Perplexity)

퍼플렉스티는 어떤 시점에서 선택할 수 있는 인식 대상 후보의 평균 개수를 의미한다. 즉, 어떤 용어의 퍼플렉스티가 2이라는 의미는 그 용어가 가질 수 있는 정답의 수가 2개라는 의미이다. 여기서는 1에 가까울수록 좋은 결과를 나타낸다. 클러스터 C_j의 퍼플렉스티 P_i은 (식 7)와 같다.

$$P_i = 2^{H(C_i)} \quad \text{(식 7)}$$

3. 실 험

이 논문에서는 전자 사전의 뜻 풀이말을 이용하여 연관된 용어끼리 자동으로 클러스터를 형성하는 실험을 하였다. 실험에 사용된 환경 변수는 [그림 4]과 같다.



[그림 4] 환경 변수

우선 자질 추출 및 선정을 사전의 뜻 풀이말을 사용하는 방법과 뜻 풀이말 자질을 확장하는 방법으로 나누었고, 추출된 자질을 *tf*idf*와 상호정보량으로 각각 달리 표현하였다. 그리고 유사도 측정을 벡터의 내적에 의한 방법과 코사인 계수를 사용한 방법으로 나누었다. 이들을 Single-link 알고리즘[12]에 적용하여 클러스터를 형성하였다. 마지막으로 클러스터링 결과를 한국어 의미 계층망과 비교하여 엔트로피와 퍼플렉스티를 계산하였다.

3.1. 대상 말뭉치

실험에 사용하기 위하여 데이터 베이스 및 XML 형식으로 되어 있는 각종 사전과 말뭉치를 가공하기 쉬운 형태의 텍스트로 변환하였다. 또한 한국어 의미 계층망과 세종 말뭉치의 태그를 일치시켜 전체 말뭉치의 일관성을 유지하였다. 각 말뭉치는 다음과 같이 구성되어있다.

- 전자 사전

전자 사전의 뜻풀이 말은 형태소 분석 태그가 부착되어 있다. <표 4>은 '학교'라는 용어의 뜻 풀이말의 구조를 나타낸다.

<표 4> '학교'의 뜻 풀이말

학교 (명) 일정_01/NNG+하/XSV+ㄴ/ETM 목적_02/NNG ^/SS 설비/NNG ^/SS 제도_01/NNG 및/MAJ 규칙/NNG+에/JKB 의거_01/NNG+하/XSV+아/EC+,/SP 교사_04/NNG+가/JKS 계속_02/NNG+적/XSN+으로/JKB 피/XPN+교육자/NNG+에게/JKB 교육/NNG+을/JKO 실시_02/NNG+하/XSV+는/ETM 기관_07/NNG+./SF

- 2002년도 세종 말뭉치

2002년도 세종 말뭉치는 21세기 세종 말뭉치 구축 프로젝트[1]의 2002년도 결과물로써 20세기 초 이후 현재까지 한국어를 대상으로 형태소 분석 및 어휘 의미 분석한 말뭉치이다. <표 5>은 2002년도 세종 말뭉치의 구조를 나타낸다.

<표 5> 세종 말뭉치의 구조

환자/NNG 들/XSN 의/JKG 얘기/NNG 를/JKO 가만히/MAG 듣_03/VV 다/EC 보/VX 면/EC 어떤/MM 때_01/NNG ㄴ/JX 너무/MAG 도/JX 기발/XR 하/XSA 아/EC 누구/NP 이/VCP 라도/EC 그런/MM 생각_01/NNG 을/JKO 하/VV 면/EC 헤어나/VV 지/EC 못하/VX ㄹ/ETM 것/NNB 이/VCP 라는/ETM 생각_01/NNG 마저/JX 들_01/VV ㄴ다/EF ./SF

- 한국어 의미 계층망

울산대 자연언어 처리 연구실에서 구축한 한국어 의미 계층 망(UWIN)[5,7]은 한국어 어휘의 계층 관계가 의미망으로 표현되어 있다. <표 6>은 한국어 의미 계층망의 구조를 나타낸다.

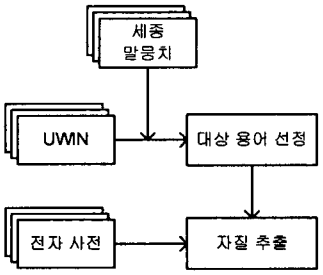
<표 6> 한국어 의미 계층망

학과_0100	과_0401
학과_0200	교육과정_0000
학과목_0000	과목_0201
학관_0201/END	학교_0000
학교_0000	교육기관_0000
학교군제도_0000	교육제도_0000

단말노드에는 /END가 표시되어있고, 의미 태그는 사전의 다의어 수준까지 구분되어 있다.

3.2. 용어 선정 및 자질 추출

이 논문에서의 자질 추출 과정은 클러스터링 대상 용어를 선정하는 단계와 선정된 용어의 자질을 전자 사전의 뜻 풀이말에서 추출하는 단계로 구성되며 [그림 5]와 같다.



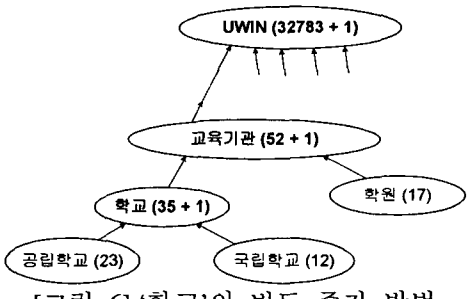
[그림 5] 자질 추출 방법

• 대상 용어 선정

이 논문에서 클러스터링 대상 용어를 선정하는 방법은 (1) 세종 말뭉치를 이용하여 UWIN의 빈도수를 측정하고 (2) 측정된 빈도수를 바탕으로 클러스터의 대표어(개념 노드)를 추출한다. 마지막으로 (3) 추출된 대표어(개념 노드)의 하위 용어 중 빈도수가 높은 100개의 용어를 추출한다.

(1) UWIN의 빈도수 측정

세종 말뭉치에서 명사를 추출하고 추출된 명사를 이용하여 한국어 의미 계층망의 용어 빈도수를 측정한다. 측정 방법은 하위 용어의 빈도수가 상위 용어의 빈도수에 반영되도록 한다. 즉, 임의의 용어는 그 자신의 빈도수에 하위 용어의 빈도수를 합한 값과 일치한다. 예를 들어 [그림 6]를 보면 용어 '학교'의 빈도수는 하위 용어 '공립학교'와 '국립학교'의 빈도수가 포함되어 있다. 이 때, '학교'라는 용어가 세종 말뭉치에서 추출한 명사 목록에 있다면 '학교'의 상위 용어인 '교육기관'에서 최상위노드인 'UWIN'까지 1씩 증가시킨다.



[그림 6] '학교'의 빈도 증가 방법

(2) 클러스터 대표어 추출

이렇게 측정된 각 노드의 빈도수를 바탕으로 용어의 빈도수가 (식 8)을 만족하는 용어를 클러스터의 대표어로 선정한다. 여기서 대표어란 하위 용어를 포함하는 개념어로써 예를 들어, [그림 6]에서 '공립학교', '국립학교', '학교', '학원'을 포함하는 대표어는 '교육기관'이다.

$$F < \alpha * f \qquad \text{(식 8)}$$

여기서 α는 임의의 상수이고, F는 최상위노드의 빈도수이고, f는 현재 노드의 빈도수이다.

이때, α의 값에 따른 클러스터 대표어 k의 개수는 <표 7>와 같다. 이 논문에서는 α를 5로 설정하여 30개의 상위노드를 대상으로 결과 클러스터와 비교하였다.

<표 7> α에 따른 클러스터 수 (대표어)

α	k	α	k	α	K	α	k
1	1	11	47	30	136	130	152
2	30	12	50	40	135	140	157
3	30	13	95	50	141	150	159
4	30	14	117	60	141	160	163
5	30	15	132	70	146	170	164
6	40	16	132	80	146	180	182
7	40	17	132	90	149	190	182
8	42	18	132	100	150	200	182
9	42	19	133	110	149		
10	47	20	134	120	152		

(3) 대표어의 하위 용어 선정

선정된 대표어의 하위 용어 중 빈도수가 높은 상위 100개의 용어를 클러스터 대상 용어로 선정한다.

• 자질 추출

클러스터링 대상 용어가 선정되고 나면 선정된 용어의 자질을 추출한다. 이 논문에서는 자질을 추출하는 방법을 (1) 사전의 뜻 풀이말을 자질로 사용하는 방법과 (2) 사전의 뜻 풀이말 자질을 확장 하는 방법(Gloss Vector)으로 나누었다.

(1) 사전의 뜻 풀이말 추출

사전의 뜻 풀이 말에서 조사, 감탄사, 기호 등과 같은 불용어를 제외한 나머지(주로 명사)를 자질로 추출한다.

(2) 자질 확장 (Gloss Vector)

이 논문에서 사용된 자질 확장 말뭉치는 2002년도 세종 말뭉치이다. 세종 말뭉치의 확장 자질을 추출하는 방법에 따라 두 가지로 구분하였다. 하나는 Bigram을 추출하는 방법이고 다른 하나는 격 관계를 추출하는 방법이다. 세종 말뭉치에서 확장에 사용될 자질을 추출할 때 대명사(NP), 수사(NR), 보조용언(VX), 지정사(VC), 관형사(MM), 부사(MA), 감탄사(IC), 조사(J), 접두사(XP), 접미사(XS), 어근(XR), 기호(S) 등은 제외했다.

▪ Bigram을 추출

세종 말뭉치에서 연속된 두 단어와 빈도수를 추출한다. 이를 바탕으로 2.1절에서 설명한 Gloss Vector에 의한 방법으로 자질을 확장한다. <표 8>은 이렇게 추출한 세종 말뭉치 Bigram 중 일부이다.

<표 8> 세종 말뭉치의 Bigram

부동산	투기_04	32
노동_03	해방	31
인과_03	관계_04	31
민족	통일_02	30
민족	해방	29
쌀	개방_02	29
위반_03	혐의	29
유엔	가입	29
노동_03	조합_01	28
일기	쓰	27
미국	대통령	26
민족	의식_02	26
사회주의	국가_01	26
차명	계좌_02	25

▪ 격 관계를 추출

세종 말뭉치에서 동사 중, '하', '되', '시키' 등과 같은 동사 파생 접미사(XSV)를 제외한 나머지 동사와의 관계가 주격과 목적격에 있는 명사를 추출하고 이를 바탕으로 2.1절에서 설명한 Gloss Vector에 의한 방법으로 자질을 확장한다. <표 9>는 이렇게 추출한 세종 말뭉치의 격 관계 테이블 중 일부이다.

<표 9> 세종 말뭉치의 격 추출

하	일_01	327
있	일_01	293
들_01	생각_01	286
있	필요	246
하	생각_01	222
하	역할	180
들리_03	소리_01	161
뜨	눈_01	157
하	생활	146
쓰	글	139
내_02	소리_01	127
하	이야기	126
감_01	눈_01	116
읽	책_01	107
하	운동_02	95
지르_03	소리_01	91
나	소리_01	91
들_01	느낌	90
자_01	잠_01	89

3.3. 사용 언어 및 알고리즘

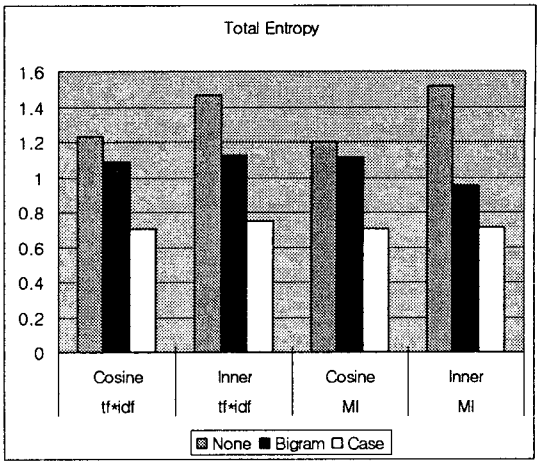
이 논문에서 사용한 알고리즘은 계층적인 클러스터링 알고리즘 중의 하나인 Single-link 알고리즘[12]을 사용하였고 펄(Perl) 언어로 구현하였다.

3.4. 평가

용어 클러스터링 시스템의 평가는 시스템에 의해 수행된 결과와 한국어 의미 계층망을 비교하여 엔트로피를 계산하였다. 그 결과를 자질확장에 따른 성능비교, 자질 표현 방법에 따른 성능 비교, 유사도 측정 계수에 따른 성능 비교로 나누어 분석하였다.

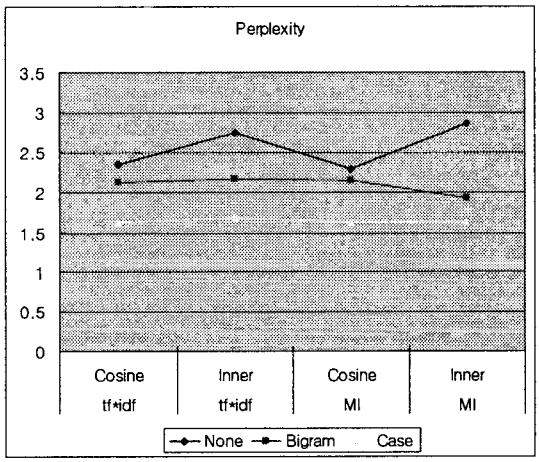
● 자질확장에 따른 성능비교

[그림 6]을 보면 용어의 뜻 풀이말만을 자질로 사용한 경우(None)보다는 자질 확장을 한 결과(Bigram, Case)가 전체적으로 성능이 좋게 나타났다. 그리고 자질 확장 시, 단순히 Bigram을 사용한 것보다는 격정보(Case)를 이용한 자질확장 방법이 성능이 우수하게 나타났다.



[그림 6] 엔트로피 측정 결과

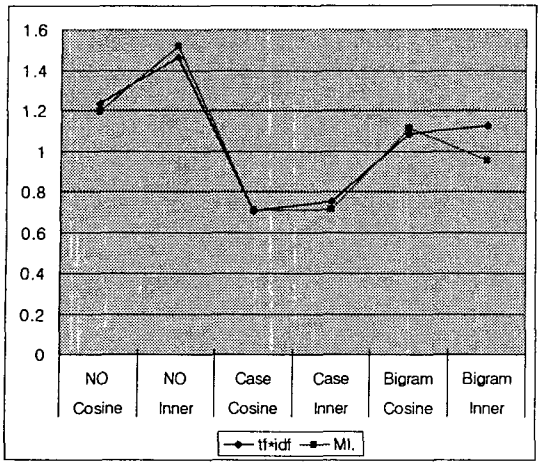
[그림 7]는 퍼플렉스티를 측정한 결과이다. 임의의 용어가 가지는 정답 클러스터 수는 격 정보를 이용한 방법이 약 1.5개로써 자질을 확장하지 않는 것보다 우수하게 나타났다.



[그림 7] 퍼플렉스티 측정 결과

● 자질 표현 방법에 따른 성능 비교

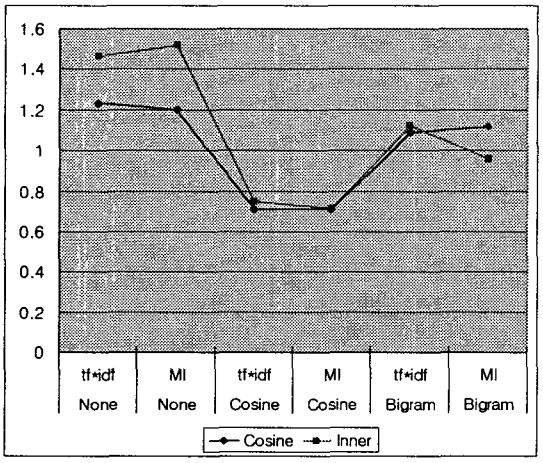
[그림 8]은 자질 표현 방법에 따른 성능 비교이다. 용어의 뜻 풀이말만을 자질로 사용하였을 경우에는 단순한 *tf*idf*에 의한 방법 보다는 상호 정보량을 사용하는 것이 성능이 좋게 나타나지만 자질확장을 한 경우에는 자질 표현 방법이 클러스터링 성능에 영향을 미치지 않는 것으로 나타났다.



[그림 8] 자질 표현 방법에 따른 성능 비교

● 유사도 측정 계수에 따른 성능 비교

뜻 풀이말만을 자질로 사용한 방법에서는 코사인 계수를 사용한 방법이 벡터의 내적에 의한 방법보다 좋게 나타났다. 그러나 자질 확장한 것에서는 두 계수의 차이가 거의 나지 않는 것으로 나타났다.



[그림 9] 유사도 계수에 따른 성능 비교

4. 결론 및 향후 과제

이 논문에서는 전자 사전의 뜻 풀이말을 이용하여 용어를 자동 분류하는 용어 클러스터링 시스템을 설계하고 구현하였다. 클러스터링 성능에 영향을 미치는 요소로 자질 선택, 자질 표현 그리고 유사도 측정 등 있는데, 이러한 요소들을 서로 다르게 설정하여 용어 클러스터링 시스템에 미치는 영향을 평가해보았다. 객관적인 성능 비교를 위하여 용어 클러스터링 결과와 한국어 의미계층망에서 추출한 클러스터를 비교하였다. 실험 결과, 용어의 뜻 풀이말만 자질로 사용한 경우에는 자질 표현과 유사도 계수를 각각 상호정보량과 코사인 계수를 사용하는 것이 좋은 성능을 보였다. 그러나 용어의 뜻 풀이말 자질을 확장한 경우에는 자질의 표현과 유사도 계수 방법의 차이는 거의 나타나지 않았다. 뜻 풀이말을 확장하는 경우에도 확장 방법이 단순한 Bigram을 이용하여 확장한 것 보다 목적어 혹은 주어 관계에 있는 격정보(Case)로 자질을 확장한 경우가 성능이 좋게 나타났다.

이 논문에서는 자질이 상대적으로 작은 용어 클러스터링의 성능에 미치는 요소를 연구함으로써 용어 클러스터링에 적합한 자질 확장방법을 소개하였다. 그래서 추후 용어 클러스터링 시스템 구축에 이러한 정보가 도움이 될 것으로 기대된다. 향후에는 좀더 다양한 클러스터링 기법을 적용하여 클러스터링 알고리즘 별로 적합한 환경 요소를 찾는 것이 필요할 것이다. 그리고 좀더 다양하고 객관적인 평가 방법을 적용하여 다양한 측면에서 클러스터링 성능을 비교함으로써 좀더 객관적인 평가가 되도록 해야 할 것이다.

참고 문헌

[1] 21세기 세종계획 전자사전 개발분과, *2000년도 연구보고서*, 문화관광부, 2000.

[2] 김건오, 고영중, 서정연, "어휘 클러스터링을 이용한 자동 문서 요약", *한국 정보 과학회 춘계 발표회 논문집*, pp. 464-465, 2002.

[3] 김영택, *자연언어처리*, 생능출판사, pp. 387-395, 2001.

[4] 김정하, 이재윤, "문헌 클러스터링 결과의 성능 평가 방법에 관한 비교 연구", *제7회 한국정보관리학회 학술대회 논문집*, pp. 45-50, 2000.

[5] 옥철영, U-WIN, *제 3 회 지식정보처리와 온톨로지 워크숍 발표자료집*, 2005.

[6] 이재윤, "단어 동시출현 기반 질의확장의 성능 최적화에 관한 연구", *연세대학교 문헌정보학과 박사학위논문*, 2003.

[7] 조평옥, "한국어 명사의 의미 계층 구조", *울산대 석사학위논문*, 1996.

[8] 한승희, 이재윤, "문헌클러스터링을 위한 유사계수간의 연관성 측정", *제6회 한국정보관리학회 학술대회 논문집*, pp. 25-28, 1999.

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

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

[11] Schütze, H., "Automatic word sense discrimination", *Computational Linguistics*, Vol. 24 No. 1 pp. 97-123, 1998.

[12] Sneath, P. H. A., Sokal, R. R., *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, Freeman. London, UK., 1973.