자질 확장에 따른 용어 클러스터링의 성능 향상

박은진[O] 김재훈 옥철영
한국해양대학교 컴퓨터공학과, 울산대학교 컴퓨터정보통신공학부
{bakeunjin[O], jhoon}@mail.hhu.ac.kr, okcy@ulsan.ac.kr

Enhancement of Word Clustering through Feature Extension

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

요 약

이 논문에서는 용어 클러스터링의 성능에 직접적인 영향을 주는 자질 확장에 따른 시스템의 성능 변화를 보았다. 객관적인 성능 비교를 위하여 용어 클러스터링 결과와 한국어 의미 계층망에서 추출한 클러스터를 비교하였다. 실험 결과, 용어의 뜻 풀이말을 자질로 사용한 경우보다 자질을 확장한 방법(Bigram, Case)이 성능이 좋게 나왔으며, 자질확장 시에 사용되는 말뭉치의 추출방법에 따라 다른 성능을 보였는데, 단순히 Bigram 정보를 사용하여 확장한 것 보다는 동사의 격 관계(Case)정보를 이용한 것이 성능이 좋게 나왔다.

1. 서 론

인터넷통계정보검색시스템[1]에 따르면 2004년 12월 기준, 우리나라 인터넷이용자수가 3,257만 명, KR도메인수가 619,446개, 국내인터넷호스트수가 5,433,591개라고 한다. 하나의 호스트에서 제공하는 정보를 가만할 때 인터넷 사용자들이 접하는 정보가 가히 엄청나다고 할 수 있다. 이렇게 방대한 정보에서 자신이 원하는 정보를 찾아내는 것은 쉬운 일이 아니다. 그래서 최근 인터넷 검색시스템은 클러스터링 기법을 이용하여 연관된 문서를 그룹화한 뒤, 사용자에게 보여줌으로써 사용자가 자신이 원하는 문서를 쉽게 찾도록 도와준다[2].

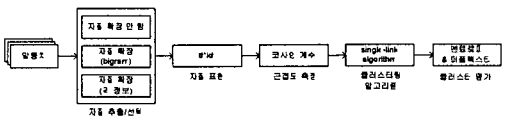
이러한 클러스터링 기법은 검색된 결과의 후처리 과정으로 사용되는 문서 클러스터링과 검색 질의 확장, 용어의 모호성 해소, 문서요약 시스템에 사용되는 용어 클러스터링으로 분류할 수 있다[2,3,4,6]. 문서 클러스터링은 연관된 문서를 하나의 클러스터로 형성하는 것을 말하며, 용어 클러스터링은 의미가 유사한 용어를 하나의 클러스터로 형성하는 것을 말한다. 일반적으로 용어 클러스터링의 경우 용어에 대한 자질로 사전의 뜻 풀이말을 이용한다. 그러나 이러한 뜻 풀이말은 문서 클러스터링의 자질에 비해 아주 작은 특징이 있다. 이러한 작은 자질로 인하여 전체 시스템의 성능이 저하되는 문제를 해결하기 위하여 자질 확장을 통해 시스템의 성능을 높이려는 연구가 활발히 진행되고 있다[3,4,6]. 이 논문에서는 자질의 크기에 민감한 용어 클러스터링 시스템에 영향을 주는 자질 확장에 따른 용어 클러스터링 성능을 평가해보았다.

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[7].

[1] http://isis.nic.or.kr
[2] http://vivisimo.com/



[그림 1] 시스템 구성

말뭉치가 주어지면 클러스터를 형성하는데 유용한 자질을 추출하게 되고, 추출된 자질은 계산이 용이한 형태인 벡터로 표현하게 되는데 일반적으로 널리 사용되는 자질 표현 방법으로는 *tf*idf*을 많이 사용한다(식 2).

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

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

이와 같이 용어-자질 행렬로 표현이 되면, 이를 다시 용어 간의 유사한 정도로 나타내는 행렬로 변환하고 유사도 측정 방법에는 코사인 계수를 이용한 방법을 많이 사용한다[8](식 3).

$$s_{ij} = Cos(w_i, w_j) = \frac{\sum_{z=1}^{m}(x_{iz} \times x_{jz})}{\sqrt{\sum_{z=1}^{m} x_{iz}^2} \times \sqrt{\sum_{z=1}^{m} x_{jz}^2}} \quad \text{(식 3)}$$

여기서 분모는 벡터의 내적이고, 분자는 각각의 자질 벡터의 거리이다.

[그림 3]과 같이 유사도 행렬을 클러스터링 알고리즘에 적용하여 연관된 용어를 하나의 클러스터로 형성한다. 일반적으로 클러스터링 알고리즘은 계층적인 방법, 분할적인 방법, 그리고 복합적인 방법으로 분류된다[7]. 이 논문에서는 계층적인 방법 중 Single-link 알고리즘[10]을 적용하였고, 펄로 구현하였다.

그리고 이 논문에서는 클러스터내의 불순한 정도를 나타내는 엔트로피(Entropy)와 임의의 용어가 가지는 클러스터의 대상후보를 나타내는 퍼플렉스티(Perplexity)를 측정하여 클러스터의 결과를 평가한다(식 4,5).

$$H(C_i) = \sum_{x \in C_i} \sum_{s \in S_j} P(x,s) \log_2 P(x,s) \quad \text{(식 4)}$$

여기서 C_i은 기계가 수행한 i번째 클러스터를 의미하고, S_j은 한국어 의미 계층망에서 추출한 j번째 클러스터를 의미한다. $P(x,s)$

는 s_i에 C_i가 속할 확률을 나타낸다. 엔트로피가 0에 가까울수록 클러스터링 결과가 우수하다는 의미로 해석된다.

$$P_i = 2^{H(C_i)} \quad \text{(식 5)}$$

3. 실 험

이 논문에서는 용어 클러스터링에 있어서 자질의 확장이 용어 클러스터링 시스템에 미치는 영향을 알아보기 위하여 [그림 1]과 같이 자질 추출/선택에서 자질을 확장하는 방법과 확장하지 않는 방법으로 나누었다. 이렇게 추출된 자질을 *tf*idf*를 사용하여 자질을 표현하였고 코사인 계수를 사용하여 용어 간의 유사한 정도를 측정하였다. 그리고 클러스터링 알고리즘은 Single-link 알고리즘[10]을 사용하였다. 마지막으로 클러스터링 결과를 한국어 의미 계층망에서 추출한 클러스터와 비교하여 엔트로피와 퍼플렉스티를 계산하였다.

3.1. 대상 말뭉치

실험에 사용된 말뭉치로는 전자 사전, 2002년도 세종 말뭉치, 그리고 한국어 의미 계층망을 사용하였다. 실험에 사용하기 위하여 데이터 베이스에 저장된 전자사전과 한국어 의미 계층망을 가공하기 쉬운 형태의 텍스트로 변환하였다. 그리고 XML 형식으로 되어 있는 세종 말뭉치에서 말뭉치만을 따로 분리해내었다. 전체 말뭉치의 일관성을 위하여 품사 태그를 같은 종류로 일치시켰다.

● 전자 사전

전자 사전의 뜻풀이 말은 형태소 분석 태그가 부착되어 있다. <표 1>은 '학교'라는 용어의 뜻 풀이말의 형태소 분석 구조를 나타낸다.

<표 1> '학교'의 뜻 풀이말의 형태소 분석 구조

학교 (명) 일정_01/NNG+하/XSV+ㄴ/ETM 목적_02/NNG ^/SS 설비/NNG ^/SS …….

● 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 2>은 2002년도 세종 말뭉치의 구조를 나타낸다.

<표 2> 세종 말뭉치의 구조

환자/NNG 들/XSN 의/JKG 얘기/NNG 를/JKO 가만히/MAG 듣_03/VV 다/EC 보/VX 면/EC …….

● 한국어 의미 계층망

울산대 자연언어 처리 연구실에서 구축한 한국어 의미 계층 망(UWIN)[5]은 한국어 어휘의 계층 관계가 의미망으로 표현되어 있다. <표 3>은 한국어 의미 계층망의 구조를 나타낸다.

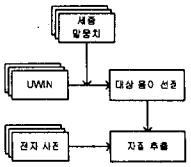
<표 3> 한국어 의미 계층망

학과_0100	과_0401
학관_0201/END	학교_0000
학교_0000	교육기관_0000

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

3.2. 용어 선정 및 자질 추출

여기서는 이 논문에서 사용된 용어 선정 및 자질 추출방법에 대해 설명한다. 이 논문에서는 한국어 의미 계층망(UWIN)에서 클러스터링 대상 용어를 선정하였고 이때 선정 기준은 세종 말뭉치에 나타난 명사에서 빈도수가 높은 용어를 우선적으로 선정하였다. 그리고 전자사전에서 선정된 용어의 자질을 추출하였다. 자질을 추출하는데 있어서 용어의 뜻 풀이말을 사용하는 방법과 뜻 풀이말을 확장하는 방법으로 구분하였다. 이러한 추출방법은 [그림 5]와 같다.



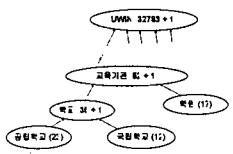
[그림 5] 자질 추출 방법

● 대상 용어 선정

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. 이때, '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.

$$F < \alpha * f \quad \text{(식 8)}$$

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

이때, α의 값에 따른 클러스터 대표어 k의 개수는 <표 4>와 같다. 이 논문에서는 α를 5로 설정하여 30개의 상위노드를 대상으로 결과 클러스터와 비교하였다.

<표 4> α에 따른 클러스터 수 (대표어)

α	k	α	k	α	k	α	k
1	1	6	40	11	47	16	132
2	30	7	40	12	50	17	132
3	30	8	42	13	95	18	132
4	30	9	42	14	117	19	133
5	30	10	47	15	132	20	134

(3) 대표어의 하위 용어 선정

선정된 대표어의 하위 용어 중 빈도수가 높은 상위 100개의 용어를 클러스터 대상 용어로 선정한다. 이렇게 선정된 용어 그룹은 클러스터일 대상 용어로 사용되고, 또한 클러스터링 결과와 엔트로피를 측정하는데 사용된다.

● 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 확장하는 방법(Gloss Vector)으로 나누었다.

(1) 사전의 뜻 풀이말 추출

사전의 뜻 풀이 말에서 조사(J), 기호(S) 등과 같은 불용어(Stop Word)를 제외한 나머지(주로 영사)를 자질로 추출한다. <표 5>은 '학교'의 뜻 풀이말을 나타내고, 이때 '학교'라는 용어의 자질로는 '일정', '목적', '설비', '제도', '규칙', '교사', '피교육자', '교육', '기관' 등이 된다.

<표 5> '학교'의 뜻 풀이말

학교 (명) 일정한 목적, 설비, 제도 및 규칙에 의거해, 교사가 계속적으로 피교육자에게 교육을 실시하는 기관.

(2) 자질 확장 (Gloss Vector)

Gloss Vector 방법은 용어의 자질을 다른 말뭉치를 이용하여 확장시키는 방법이다[9]. 이 논문에서 사용된 자질 확장 말뭉치는 2002년도 세종 말뭉치이다. 여기서 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.

세종 말뭉치에서 Bigram을 추출하는 방법은 연속된 두 명사를 추출하였다. <표 6>는 '교육'과 연속된 두 명사를 추출한 결과 중 일부를 나타낸다. 그리고 세종 말뭉치에서 격 관계를 추출하는 방법은 동사 중에서 '하', '되', '시키' 등과 같은 동사 파생 접미사(XSV)를 제외한 나머지 동사와의 관계가 주격과 목적격에 있는 명사를 추출한다. <표 7>은 '교육'을 주격 혹은 목적격 관계에 있는 동사 중 일부를 나타낸다.

<표 6> '교육'을 포함하는 Bigram

교육	학교	36
교육	민족	15
교육	식민지	15
교육	위하	15
교육	나라	14
교육	자녀	12

<표 7> '교육'을 포함하는 격 관계

교육	하	21
교육	위하_01	14
교육	통하	12
교육	시키	11
교육	이루_01	8
교육	바라보	2

이렇게 확장에 사용될 자질을 추출하고 나면 용어의 뜻 풀이말을 확장하게 된다. 예를 들어 '학교'라는 용어 뜻 풀이말의 자질 중에서 '교육'이라는 자질의 확장 과정을 살펴보면, '교육'이라는 자질은 <표 6>에서 '교육'과 같이 나타난 단어 '학교', '민족', '식민지', '위하', '나라' 등으로 대치된다. 이때 '교육'에 대한 자질 벡터가 <표 8>와 같이 다시 계산된다.

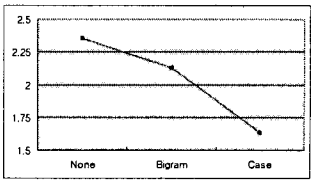
<표 8> '학교' 뜻 풀이말의 Gloss Vector

뜻풀이말 \ 공기어	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**

같은 방법으로 뜻 풀이말 '목적', '설비', '제도', '규칙', '피교육자', '교육', '기관' 등이 확장하게 되고, 결국 '학교'의 뜻 풀이말의 확장 벡터(Gloss Vector)는 <표 8>와 같다.

3.3. 평가

용어 클러스터링 시스템의 평가는 시스템에 의해 수행된 결과와 한국어 의미 계층망에서 추출한 결과를 비교하여 엔트로피를 계산하였다. [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 격 정보(Case)를 이용한 자질확장 방법이 성능이 우수하게 나타났다.



[그림 6] 엔트로피와 퍼플렉스티 측정 결과

4. 결론 및 향후 과제

이 논문에서는 자질이 작은 용어 클러스터링의 성능에 직접적인 영향을 주는 자질 확장에 따른 시스템의 성능 변화를 보았다. 객관적인 성능 비교를 위하여 용어 클러스터링 결과와 한국어 의미 계층망에서 추출한 클러스터를 비교하였다. 실험 결과, 용어의 뜻 풀이말을 자질로 사용한 경우보다 자질을 확장한 방법(Bigram, Case)이 성능이 좋게 나왔다. 자질확장 시에 사용되는 말뭉치의 추출방법에 따라 다른 성능을 보였는데, 단순히 Bigram 정보를 사용하여 확장한 것 보다는 동사의 격 관계(Case)정보를 이용한 것이 성능이 좋게 나왔다.

이 논문에서는 자질이 작은 용어 클러스터링에 요구되는 자질 확장에 따른 시스템 성능 변화를 연구함으로써 용어 클러스터링에 중요한 자질 확장방법을 소개하였다. 추후 용어 클러스터링 시스템 구축에 이러한 연구가 도움이 될 것으로 기대된다. 향후에는 좀더 다양한 클러스터링 기법을 적용하여 클러스터링 알고리즘별로 적합한 자질확장 방법을 찾는 것이 필요할 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] 김명철, "공기기반 용어간 유사도를 이용한 정보검색 질의확장 비교연구", *한국과학기술원 전산학과, 박사학위 논문*, 1999.

[4] 이상훈, 김기태 , "클러스터링 기법을 이용한 키워드 유사도 순위화 알고리즘에 따른 사용자 질의 확장", *한국 정보 과학회 2003년 춘계학술대회* , 2003.

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

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

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

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