퍼지 추론을 이용한 소수 문서의 대표 키워드 추출에 대한 유용성 평가

노순억[0] 김병만 신윤식 임은기
(sorho[0], bmkim, ysshin, eklim)@se.kumoh.ac.kr

Evaluation on the usefulness of Representative Keyword Extraction from Few Documents through Fuzzy Inference

Sun Ok Rho[0] Byeong Man Kim Yoon Sik Shin En Ki Lim
Dept. of Software Engineering, Kumoh National University of Technology

요 약

본 논문은 퍼지 추론을 이용하여 소수문서로부터의 대표 용어들을 추출하고 가중치를 부여한 기존 방법의 유용성을 평가하고자 GIS (Generalized Instance Set) 알고리즘에 이를 적용시켜 보았다. GIS 는 학습 문서 집합에 대한 클러스터링 과정을 통해 문서 그룹들을 생성하고 이들에 대한 선형 분류기들을 유도한 뒤 k-NN 알고리즘을 적용하는 방법이다. GIS의 일반화(generalization) 과정에 Rocchio, Widrow-Hoff 및 퍼지 추론을 이용한 방법을 적용시켜 문서 분류 성능을 비교하였다. 긍정적 문서 집합에 대한 실험에서 비교적 우수한 성능 향상을 보여줌으로써 퍼지 추론을 이용한 방법의 유용성을 확인 할 수 있었다

1. 서 론

문서 검색 또는 문서 필터링 시스템들은 보통 새로운 문서가 사용자의 관심사(interest) 또는 특정 범주(class)에 대해 얼마나 적합한지 부적합한지를 판단하기 위해서 자동화된 문서 분류기들에 의존한다[1][2]. 이들 분류기들은 사용자 혹은 외부로부터 정의해 놓은 학습문서 집합을 이용하여 학습과정을 거치면서 분류체계들을 유도해 낸다. 이들 분류기들에는 각자의 장.단점을 내포하고 있다. 예로 들자면 Rocchio 와 Widrow-Hoff 와 같은 선형 분류기들은 벡터 공간 모델에 기반 하여 문서를 표현하는 벡터의 가중치 계산 시와 학습 문서집합의 중심 벡터의 가중치 계산 시에 용어의 발생 빈도수(TF)나 역문헌 빈도수(IDF) 정보를 이용한다[1]. 이는 문서 내 또는 문서 집합 내 용어들 사이의 관련성을 용어의 가중치 계산에 반영하고 있지 않음을 의미한다. 일반적으로 문서는 특정 주제를 표현하는 용어들을 포함하고 있으며 이들 용어들 중에는 해당 주제에 가장 밀접한 전문적인 용어들이 존재할 수 있다. 이들 용어들간에는 유사한 동시 발생 빈도(co-occurrence frequency) 성향을 가질 수 있다. 다시 말해 어떤 중요 용어에 수반하여 관련 있는 다른 용어가 함께 나타날 가능성이 높고 이러한 용어들의 중요성은 해당 용어의 TF나 IDF와 같은 통계적인 수치만으로 계산하기 보다는 앞에서 언급한 바와 같은 용어들의 관련성을 계산에 반영하는 것이 보다 효과적일 것이다. 이러한 접근 방법으로써 퍼지 추론으로 문서집합의 핵심적인 대표 용어들을 추출하고 이들과의 동시 발생 빈도수의 유사성을 가중치 계산에 반영하는 방법이 제시되었다[3].

본 논문에서는 위 방법이 지닌 학습문서 집합 크기에 대한 제약성을 극복하고자 GIS 알고리즘[4]에 위 방법을 적용하여 다수의 학습문서를 가진 범주들에 대한 문서 분류 실험을 통해 성능 향상의 가능성을 확인하고자 한다.

2. 관련연구

문서 분류에 사용된 대표적인 분류기들에는 Decision tree, Dicision rule, Neural network, Rocchio, Widrow-Hoff, k-NN, GIS, SVM 등 이 있으며 이들 중에서 문서 집합에서 대표 용어를 추출하고 가중치를 부여하는 문제[3]와 유사한 성격을 지닌 학습 문서 집합의 중심 벡터를 구성하는 분류 방법들로는 Rocchio 와 Widrow-Hoff 그리고 GIS(Generalized Instance Set) 방법들이 있다[1][4][2].

Rocchio 분류 방법은 벡터 공간 모델에서 적합성 피드백을 위한 Rocchio 식을 문서 분류에 적용한 것으로 아래 식-1 를 사용하여 배치모드(batch mode)로 문서 집합의 중심 벡터를 계산한다[1].

$$w_j = \alpha w_{1,j} + \beta \frac{\sum_{i \in C} x_{i,j}}{n_C} - \gamma \frac{\sum_{i \notin C} x_{i,j}}{n - n_C} \qquad \text{(식-1)}$$

아래 식-2 는 Widrow-Hoff 분류 방법(식-2)으로써 한 단계마다 하나씩 문서를 처리하면서 중심 벡터의 가중치들을 갱신하고 있으며 온라인(on-line) 환경에서 사용될 수 있다[1].

$$w_{i+1,j} = w_{i,j} - 2\eta(w_i \bullet x_i - y_i)x_{i,j} \qquad \text{(식-2)}$$

퍼지 추론을 이용한 소수 문서의 대표 키워드 추출(이하 RKEF)[3]에서는 사용자가 제시한 소수의 예제 문서집합으로부터 사용자의 관심사항을 가장 잘 대변하는 대표 용어들을 추출하고 이들의 가중치를 부여하는 문제를 다루고 있다. 이는 앞에서 언급한 Rocchio와 Widrow-Hoff 방법들이 학습 문서 집합을 대표하는 중심 벡터를 구성하는 것과 성격이 유사함으로 문서분류 문제에 위 방법을 사용할 수 있다(식-4 참조). Reuters-21578 내의 소수의 긍정적 학습문서 집합을 가진 범주들에 대한 실험 결과로 Rocchio 와 Widrow-Hoff 보다 나은 성능을 보여주고 있다[3]. 그러나 많은 수의 긍정적 학

* 본 연구는 한국과학재단 목적기초연구(2000-1-51200-008-2) 지원으로 수행되었음.

습문서(예제 문서)를 가진 범주에 대해서 위 방법을 적용하기에는 어려움이 있을 것으로 여겨진다. 긍정적 학습문서 수가 많아 질 수록 추출되는 핵심적인 초기 대표 용어들의 수가 증가 할 가능성이 높아지며 이러한 경우 이들을 기준으로 계산된 용어발생 빈도수 유사도 정보는 그 유용성이 낮아질 수 있다(식-5 참조). 따라서 다수의 긍정적 문서 집합을 소수의 문서 집합들로 분할 할 수 있는 방법 즉 클러스터링 방법을 이용한다면 많은 학습문서를 가진 범주에 대한 분류 문제에서도 효과적일 수 있다. 그림 1 은 [3] 에서 제시한 소수 예제 문서 집합의 대표 벡터를 구성하는 과정을 보여주고 있으며 크게 3 단계로 다음과 같이 요약 할 수 있다.

(1) 퍼지 추론을 이용한 대표 용어 중요도 계산
(2) 초기 대표 용어 선택
(3) 용어 가중치 재산정과 대표용어 자동확장

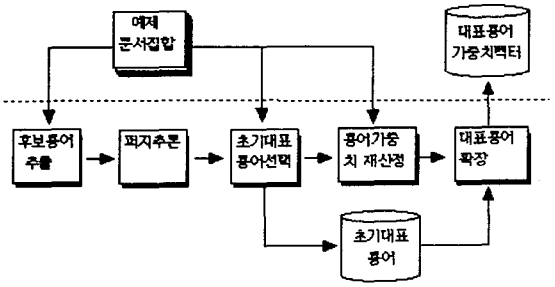

그림 1. 예제 문서집합에 대한 대표 벡터의 생성 과정

GIS(Generalized Instance Set) 방법은 선형 분류기와 k-NN 알고리즘의 장.단점을 고려하면서 각각의 방법을 결합한 방법이다[4]. GIS 알고리즘을 통해 학습 문서들을 클러스터링하고 generalized instance 들을 구축한 다음 이들을 사용한다. 그림 2 은 GIS 알고리즘의 일부분을 보여주고 있다. 그림 2 의 일반화 함수(Generalization function)에서 Rocchio, Widrow-Hoff 그리고 RKEF 방법을 적용시킬 수 있다.

```
7)  Repeat
8)      G' := G.
9)      G := Generalise(G', k).
10)     Rank instances in T according to the similarity metric with G.
11)     Compute Rep(G).
12) Until Rep(G) < Rep(G')
```

그림 2. GIS 알고리즘의 generalized instance 탐색 부분

3. 학습문서 집합의 대표 벡터 구성

아래 식-3 은 퍼지 추론에 사용된 퍼지 입력 변수들의 입력값을 계산하기 위한 식들을 나타내고 그림 3와 그림 4 은 퍼지 입.출력 변수들과 추론에 사용된 규칙들을 각각 나타내고 있다(퍼지 추론 부분에 관한 자세한 설명은 [3] 참조).

$$NTF_i = \frac{TF_i}{DF_i} \div \max_j\left[\frac{TF_j}{DF_j}\right],\quad NDF_i = \frac{DF_i}{TD} \div \max_j\left[\frac{DF_j}{TD}\right],$$

$$NIDF_i = NIDF_i \div \max_j[NIDF_j] \qquad \text{(식-3)}$$

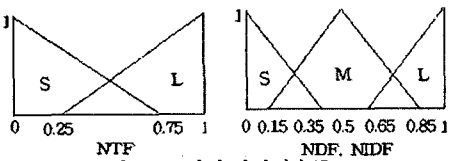

그림 3-(a). 퍼지 입력변수들

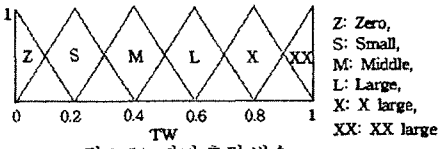

그림 3-(b). 퍼지 출력 변수

NIDF \ NDF	S	M	L
S	Z	Z	S
M	Z	M	L
L	S	L	X

NTF = S

NIDF \ NDF	S	M	L
S	Z	S	M
M	S	L	X
L	S	X	XX

NTF = L

그림 4. 퍼지 추론 규칙

식-4 은 RKEF 에서 사용된 계산식으로써 사용자가 제시한 소수 예제 문서집합에 대해서 이를 대표하는 벡터를 구성하는 최종 계산식을 보여주고 있다[3]. 식-5 는 퍼지 추론 과정에 의해 선택된 핵심적인 초기 대표 용어들과 다른 일반 용어들간의 발생 빈도수의 유사성을 계산하기 위해 사용된다. 핵심 용어들과의 발생 빈도수 유사 정도에 따라 용어의 가중치 값을 조정하게 하는 역할을 한다.

$$w_j = \sum_{i \in C}(x_{i,j} \times RD_{i,j}(K)) \qquad \text{(식-4)}$$

$$RD_{i,j}(K) = 1 - \log_p\left(\sqrt{\frac{\sum_{k=1}^{n}(TF_{i,k} - TF_{i,j})^2}{n}}\right) \qquad \text{(식-5)}$$

w_j : 대표 벡터에서 용어 j 의 가중치
$x_{i,j}$: 문서 i 에서 용어 j 의 가중치
K : 퍼지 추론으로 선택된 초기 대표 용어 집합
$RD_{i,j}(K)$: 문서 i 에서 용어 j 와 K 간의 관련정도
$TF_{i,k}$: 문서 i 에서 초기 대표 용어 k 의 발생 빈도수
$TF_{i,j}$: 문서 i 에서 용어 j 의 발생 빈도수
n : 초기 대표 용어 수

식-6 는 학습 문서 집합 중에서 부정적 문서집합을 해당 범주의 대표 벡터를 구성하는데 사용하기 위해서 본 논문에서 사용된 계산식이다. 핵심적인 초기 대표 용어 집합을 퍼지 추론을 통해서 긍정적 문서 집합과 부정적 문서 집합으로부터 개별적으로 추출하여 용어 가중치 재산정에 사용하였다. GIS 알고리즘에 사용되어 질 경우 적용 대상 학습문서 집합은 G 와의 유사도 계산 후 rank 된 상위 k 개의 문서들이다(그림 2 참조).

$$w_j = \alpha \frac{\sum_{i \in C}(x_{i,j} \times RD_{i,j}(K_p))}{n_C} - \beta \frac{\sum_{i \notin C}(x_{i,j} \times RD_{i,j}(K_n))}{n - n_C} \qquad \text{(식-6)}$$

w_j : 대표 벡터에서 용어 j 의 가중치
$x_{i,j}$: 문서 i 에서 용어 j 의 가중치
$RD_{i,j}(K_p)$: 문서 i 에서 용어 j 와 긍정적 문서 집합에서 선택된 초기 대표용어 집합 K_p 간의 관련정도
$RD_{i,j}(K_n)$: 문서 i 에서 용어 j 와 부정적 문서 집합에서 선택된 초기 대표용어 집합 K_n 간의 관련정도
n : 클래스 C의 학습문서 수

4. 실험 및 결과

실험 문서 집합으로 Reuters-21578 을 사용했다. Reuters-21578 의 TOPICS 범주들을 선택하였으며 ModApte 버전을 사용했고 라벨이 없는 문서들은 제외시켰다. 실험 대상으로 테스트 문서 집합

과 학습 문서 집합에 적어도 하나의 문서를 각각 포함하고 있는 90 개의 범주 중에서 다수의 학습 문서들을 가진 상위 20개의 범주들을 선택하였다. 태스트 문서 집합은 3019개의 문서들을 포함하고 있다. 용어의 역문헌 빈도수(IDF)값을 구하기 위해 90개의 범주들에 속하는 7770개의 학습 문서 집합으로부터 문서 빈도수 정보를 구하였다.문서를 표현하는 특징 벡터들의 가중치는 TF x IDF 로 계산하였다. 유사도 계산식은 cosine 식을 사용했고 성능 측정 방법으로 11-point average precision 을 사용하였다.

실험시 사용된 조정 상수(parameter)들의 설정값들은, 부정적 문서들을 포함한 실험에서 Rocchio의 경우 α=0, β=1, γ=1, Widrow-Hoff 의 경우 η=0.25 , 식-6 의 경우 α=1, β=1 로 두었으며 긍정적 문서들의 정보만을 이용한 실험에서 Rocchio의 경우 α=0, β=1, γ=0, Widrow-Hoff의 경우 η=0.25, 식-6 의 경우 α=1, β=0 으로 두었다. 퍼지 추론을 이용한 방법에서 식-5의 p 는 10 으로 각각 동일하게 사용하였다. GIS 알고리즘의 경우 일반화 함수(Generalization function)에 사용된 k 값(그림 2 참조)은 10 에서 150 사이의 10 단위로 선택한 총 15개의 k 값들에 대해 분류 실험을 수행하였다.

표 1 은 GIS 알고리즘의 일반화 함수에서 부정적 문서들을 사용했을 경우, 위의 k 값들에 대한 실험 결과들 중에서 일반화 과정에 사용된 방법별로 각 범주별 가장 좋은 성능값들을 선택해서 보여주고 있다. 아울러 표 2 는 부정적 문서들을 제외한 긍정적 문서들만을 사용했을 경우 선택된 성능값들을 보여주고 있다.

표 1. 긍정적 문서 및 부정적 문서를 일반화에 모두 사용한 경우 20 개의 범주들에 대한 성능

Category	RO	WH	Best		
			GIS-R	GIS-W	GIS-F
nat-gas	0.567	0.633	0.723	0.694	0.667
soybean	0.576	0.758	0.78	0.74	0.779
veg-oil	0.568	0.703	0.739	0.716	0.701
gold	0.833	0.8	0.862	0.866	0.853
gnp	0.743	0.914	0.932	0.915	0.93
coffee	0.929	0.964	0.988	0.991	0.985
oilseed	0.553	0.723	0.663	0.665	0.613
sugar	0.694	0.889	0.91	0.914	0.917
dlr	0.682	0.705	0.805	0.777	0.787
money-supply	0.412	0.676	0.726	0.725	0.727
corn	0.661	0.821	0.898	0.857	0.901
ship	0.865	0.876	0.88	0.866	0.876
wheat	0.732	0.845	0.893	0.874	0.929
interest	0.695	0.718	0.803	0.79	0.802
trade	0.735	0.769	0.788	0.77	0.81
crude	0.804	0.857	0.88	0.838	0.892
grain	0.779	0.926	0.937	0.906	0.946
money-fx	0.581	0.749	0.694	0.686	0.688
acq	0.885	0.911	0.877	0.822	0.875
earn	0.953	0.954	0.967	0.966	0.964
Average	0.712	0.810	0.837	0.819	0.832

표 1 실험 결과를 살펴보면 GIS-W(GIS+Widrow-Hoff) 가 GIS-R(GIS+Rocchio) 보다 성능이 낮고 GIS-R와 GIS-W의 성능과 RO (Rocchio)와 WH(Widrow-Hoff)의 성능사이에 큰 차이를 보여주지 않고 있다. 이것은 방법별로 최적의 k 값을 찾아서 사용하지 않았기 때문이다. 그러나 이러한 다소 제약된 실험환경에서도 퍼지 추론을 이용한 방법의 유용성을 충분히 확인할 수 있을 것으로 판단되어 방법별 최적의 k 값들을 찾는 실험은 제외하였다.

퍼지 추론을 이용한 방법을 GIS에 적용한 결과(GIS-F)는 Rocchio 를 GIS 에 적용한 결과(GIS-R)에 비해서 별다른 성능 향상을 보여주지 않았다. 그러나 표 2 는 GIS 알고리즘의 일반화 함수에서 부정적 문서들을 제외한 긍정적 문서들만을 사용했을 경우 다른 비교 방법들에 비해서 향상된 성능을 보여주고 있다. 따라서 부정적 문서들을 고려한 퍼지 추론 방법의 개선을 통해 성능 향상을 기대 할 수 있음을 확인 할 수 있었다.

표 2. 긍정적 문서만을 일반화에 사용한 경우 20 개의 범주들에 대한 성능

Category	RO-P	WH-P	Best		
			GIS-RP	GIS-WP	GIS-FP
nat-gas	0.492	0.494	0.518	0.599	0.643
soybean	0.639	0.589	0.642	0.654	0.738
veg-oil	0.626	0.63	0.657	0.651	0.756
gold	0.854	0.843	0.861	0.863	0.846
gnp	0.815	0.82	0.831	0.835	0.871
coffee	0.936	0.979	0.947	0.936	0.989
oilseed	0.482	0.425	0.497	0.508	0.601
sugar	0.738	0.776	0.793	0.807	0.882
dlr	0.635	0.686	0.719	0.726	0.751
money-supply	0.334	0.587	0.624	0.607	0.726
corn	0.644	0.624	0.658	0.654	0.797
ship	0.821	0.745	0.831	0.821	0.854
wheat	0.764	0.798	0.808	0.803	0.861
interest	0.636	0.72	0.731	0.738	0.793
trade	0.717	0.661	0.733	0.74	0.749
crude	0.778	0.801	0.809	0.808	0.846
grain	0.802	0.871	0.859	0.866	0.867
money-fx	0.582	0.537	0.616	0.615	0.663
acq	0.576	0.727	0.707	0.708	0.792
earn	0.961	0.947	0.963	0.962	0.962
Average	0.692	0.713	0.740	0.745	0.799

5. 결 론

본 논문에서는 소수의 긍정적 문서 집합을 대상으로 문서들의 내용을 대표하는 중요 용어들을 추출하고 이들의 가중치를 부여하는 문제를 해결하기 위한 방법인 퍼지 추론 및 용어 발생 빈도수의 유사성을 이용한 가중치 재산정 접근 방법을 GIS 알고리즘에 적용시켜 문서분류 성능을 비교해 보았다. GIS 알고리즘에 적용시켜 봄으로써 소수 학습 문서 집합을 대상으로 한다는 제약성을 극복할 수 있었으며 긍정적 문서들만을 일반화에 사용한 실험에서 나은 성능을 보여줌으로써 성능 향상의 가능성을 확인할 수 있었다.

향후 부정적 문서 집합을 고려한 퍼지 추론 방법의 개선에 관한 연구를 진행 할 것이다.

참고 문헌

[1] D.D.Lewis, R.E.Schapore, J.P.Call, and R.Papka. "Training algorithms for linear text classifiers", In Proceedings of the Nineteenth International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 298-306, 1996.

[2] Sebastiani, F." Machine Learning in Automated Text Categorisation" , Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell' Informazione, 1999.

[3] 노순억, 김병만, 허남철, "퍼지 추론을 이용한 소수 문서의 대표 키워드 추출", 한국퍼지 및 지능시스템학회 논문지, vol.11, No.9, pp.837-843, 2001.

[4] W.Lam, C.Y.Ho. "Using a Generalized Instance Set For Automatic Text Categorization", ACM SIGIR Conference, 1998.