K-NN과 객체 지향 시소러스를 이용한 웹 문서 자동 분류

방선이[0] 양재동
전북대학교 전산통계학과
(sibang, jdyang)@cs.chonbuk.ac.kr

Automatic Document Categorization Using K-Nearest Neighbor Algorithm and Object-Oriented Thesaurus

Sun-Lee Bang[0] Jae-Dong Yang
Dept. of Computer Science, Chonbuk National University

요 약

문서 자동 분류에는 통계적인 기법과 machine learning 기법의 많은 알고리즘들이 이용되고 있다. 통계적인 기법 알고리즘을 이용한 문서 분류는 높은 성능을 보이지만 분류할 카테고리가 둘 이상인 경우가 빈번할 경우에는 정확률이 급격히 저하되는 단점이 있다.

본 논문에서는 K-NN알고리즘을 이용하여 일차적인 문서 분류를 수행한 후 특정 카테고리로 분류하기에 애매모호한 경우가 생길 경우 시소러스의 일반화 관계와 연관화 관계를 이용하여 모호성을 줄임으로써 문서 자동 분류의 성능을 높이기 위한 새 기법을 제안한다.

1. 서 론

최근, WWW의 발전으로 인하여 인터넷의 사용과 정보의 양이 급증함에 따라 문서 형식의 데이터가 웹상에 많이 존재하고 있다. 방대한 문서 데이터로부터 유용한 정보를 효과적으로 획득하기 위해서 문서 분류에 대한 필요성 또한 증가하고 있다. 문서 분류란, 문서의 내용을 파악하여 미리 정의되어 있는 카테고리 중 어느 카테고리에 속하는가를 결정하는 것이다. 전문가를 통한 문서분류는 정확성이 높지만 많은 시간과 노력, 비용이 들게 되므로 문서 자동 분류에 대한 연구가 계속 진행되고 있다. 대표적인 문서 분류 알고리즘으로는 베이지안 확률분류(Bayesian classifier), 결정 트리(decision tree), 최근접 이웃분류(K-nearest neighbor classification), 규칙 학습(rule learning algorithm), 신경망(neural networks), 퍼지개념을 이용한 알고리즘 등이 있다[1,2]. 이러한 문서 분류 알고리즘은 80퍼센트 정도의 정확성을 보이고 있지만, 특정 카테고리로 문서를 분류하는데 애매모호한 경우의 발생이 빈번하기 때문에 그 이상의 정확도를 얻는 데에는 한계를 보이고 있다[3].

본 논문에서는 기존의 여러 분류 알고리즘 중 가장 간단하면서도 좋은 성능을 지닌 K-NN 알고리즘의 정확도를 유지하면서 객체 지향 시소러스를 이용하여 문서 자동 분류에 대한 성능을 높이기 위한 새로운 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 K-NN 알고리즘에 대해서 살펴보고 K-NN을 이용한 문서 분류 과정을 기술한다. 3장에서는 객체 지향 시소러스를 사용하여 K-NN을 이용한 문서 분류에서 생기는 모호성을 해결하는 방법을 기술하고 마지막으로 4장에서는 결론 및 향후 과제를 제시한다.

2. K-NN 알고리즘

K-NN 알고리즘은 전문가에 의해 이미 분류되어 있는 문서들로부터 단어들의 출현 빈도에 대한 정보를 추출하여 문서벡터를 형성하고 새로운 문서에 대한 문서 벡터와의 유사도를 따져 이웃하는 순위를 정한 후, K개의 가장 가까운 문서를 이용하여 새로운 문서가 속할 카테고리를 예측한다[1,3]. 각 문서들에 대한 문서벡터와 새로운 문서에 대한 문서벡터 사이의 유사도는 유클리디안 거리나 두 벡터 사이의 각도를 나타내는 코사인 계수 등에 의해 구해지고 K개의 이웃하는 문서는 각 문서와 새로운 문서 간의 유사도에 의해 중요도가 맺어진다. K개의 이웃하는 문서가 정해지면 이들이 속하는 카테고리를 따져 중요도가 가장 높은 하나의 카테고리로 문서를 분류하므로 모든 문서를 분류할 수 있다.

K-NN을 이용한 문서분류는 다음과 같다.

전체 카테고리가 $\forall = \{c_1, c_2, \cdots, c_n\}$ 이고 전체 문서 집합은 D 일 때, 새로운 문서 d 에 대해 K개의 가까운 문서 $K-NN(d)$ 가 K-NN 알고리즘에 의해 다음과 같이 얻어졌다고 가정하자.

$$K-NN(d)=\{d_j \in D \mid 1 \le j \le K\}$$

[정의 1] $K-NN(d)$ 의 각각의 문서 d_j 와 카테고리 c_i $(i=1,2,\cdots,n)$에 대해 $c_i(d_j)$ 는 다음과 같이 정의된다.

d_j 가 카테고리 c_i , $1 \le i \le n$ 에 속하면 $c_i(d_j)=1$ 이고,

그 외의 경우, $c_i(d_j)=0$ 이다.

이 때, 각 카테고리 $c_i\,(i=1,2,\cdots,n)$에 문서 d 가 속할 관련 정도 w_i 는 $K-NN(d)$ 를 이용하여 구해보면 다음과 같다.

$$w_i = \sum_{j=1}^{K} c_i(d_j) \quad (i=1,2,\cdots,n)$$

[정의 2] 문서 d 가 속할 관련정도 w_i 와 함께 $c_i(d,w_i)$ 로 나타낼 수 있다.

[정의 3] 문서가 속할 관련 정도 w_i 에 의해 후보 카테고리 집합은 다음과 같이 정의된다.

$$C(d)=\{c_i(d,w_i) \mid w_i \ge Threshold\ ,\ i=1,2,\cdots,n\}$$

후보카테고리가 하나인 경우는 문서가 c_i 카테고리에 속하게 된다. 그러나 두 개 이상인 경우는 K-NN에서는 관련 정도가 가장 높은 카테고리를 문서 d 가 속할 카테고리로 정해버림으로써 관련정도가 같거나 약간 작은 다른 카테고리들을 무시하게 되므로

정확도를 떨어뜨리는 단점이 있다. 다음 절에서는 시소러스의 일반화 관계와 집성화/연관화 관계를 이용함으로써 문서의 카테고리 할당에서의 모호성을 줄여주는 기법에 대해 기술하도록 한다.

3. K-NN과 시소러스를 이용한 문서 분류

3.1 카테고리의 구조

[정의 4] 시소러스의 레벨 수를 n이라 하면 시소러스 구조를 반영하여 설정되는 전체 카테고리 집합 $\forall$는 다음과 같이 정의된다.

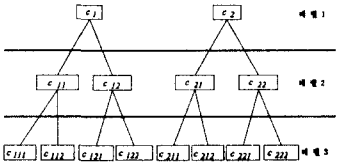
$$\forall = \{c_{i_1 i_2 \cdots i_k} \mid i_k \neq 0 \in I^+ ,\ k = 1,2,\cdots,n\}$$

$c_{i_1 i_2 \cdots i_k}$, $1 \le k \le n$ 에 대해,

c_{i_1} 은 시소러스에서의 i_1 번째의 최상위 개념이고,

$c_{i_1 i_2 \cdots i_s}$, $2 \le s < k$ 은 $s-1$ 단계에서의 i_{s-1} 번째 개념의 i_s 번째 하위 개념이며, $c_{i_1 i_2 \cdots i_k}$ 은 $k-1$ 단계에서의 i_{k-1} 번째 개념의 i_k 번째 인스턴스 이다.

시소러스 구조에 따른 카테고리 구조는 <그림 1>과 같이 도식화 할 수 있다.



<그림 1> 시소러스 구조에 따른 카테고리 구조의 예

$c_{i_1 i_2 \cdots i_k} \in \forall$에 대해 다음과 같은 정의를 내릴 수 있다.

[정의 5]

$c_{i_1 i_2 \cdots i_k}$ 에 대해 $k+1$ 레벨에서의 개념이 존재하지 않을 때, $c_{i_1 i_2 \cdots i_k}$ 는 최하위 레벨의 개념이다.

[정의 6]

문서 d 가 레벨 $k \le n$ 중 어떤 카테고리에 속한다면, 시소러스의 일반화계층에 따라 그 카테고리의 상위개념에 해당하는 카테고리에도 속한다. 즉, d 가 $c_{i_1 i_2 \cdots i_k}(d) = 1$ 이면 $c_{i_1 i_2 \cdots i_{k-1}}(d) = 1$ 이다.

전체 카테고리 중 문서 d 가 속할 수 있는 카테고리가 c 일 때, $c(d)$ 는 다음과 같은 특성을 가진다.

[명제 1] 문서 d 가 $c_{i_1 i_2 \cdots i_k}$ 카테고리에 속하면,

$c_{i_1 i_2 \cdots i_s}(d)=1$ $s=1,2,\cdots,k$, $1 \le k \le n$ 이고,

$c_{i_1 i_2 \cdots i_k}$ 카테고리에 속하지 않으면 , $c_{i_1 i_2 \cdots i_k}(d)=0$ 이다.

증명) 생략

[명제 2] $w_{i_1 i_2 \cdots i_s} \ge w_{i_1 i_2 \cdots i_k}$, $s=1,2,\cdots,k-1$, $1 \le k \le n$ 이다.

증명)생략

[명제 3] 후보 카테고리 집합을 일반화하면.

$C(d)=\{c_{i_1 i_2 \cdots i_s} \in \forall,\ s=1,2,\cdots,k \mid c_{i_1 i_2 \cdots i_k}(d,w),\ w \ge Threshold\}$ 이다.

증명)생략

3.2 시소러스에 의한 K-NN 카테고리할당의 모호성 제거

[정의 7] 카테고리 $c_{i_1 i_2 \cdots i_k}$ 에 대한 상위 카테고리를 다음과 같이 정의한다.

$$Sup(c_{i_1 i_2 \cdots i_k}) = c_{i_1 i_2 \cdots i_{k-1}}$$

[정의 8] 후보 카테고리집합 $C(d)$ 에 대해 축약된 후보 카테고리 집합을 $C_R(d)$ 라 정의한다.

$$C_R(d) = C(d) - \{Sup(c_{i_1 i_2 \cdots i_k}) \mid c_{i_1 i_2 \cdots i_k} \in C(d)\}$$

$C_R(d)$ 가 나오는 형식은 크게 $C_R(d)$ 의 원소 $c_{i_1 i_2 \cdots i_k}$ 에 대해

- $|C_R(d)|=1$
- $|C_R(d)| \ge 2$ 이면서 i_1 (최상위레벨의 개념)이 동일한 경우
- $|C_R(d)| \ge 2$ 이면서 i_1 (최상위레벨의 개념)이 상이한 경우

로 나눌 수 있다.

$|C_R(d)|$ 가 1인 경우는 최하위 레벨에 해당하는 개념이 하나인 경우이고, $|C_R(d)|$ 가 2 이상인 경우는 최하위 레벨에 해당하는 개념이 두 개 이상인 이다.

3.2.1. $|C_R(d)|$ 가 1인 경우

문서 d 가 속할 카테고리를 명확하게 결정할 수 있으므로 문서 d 는 $C_R(d)$ 집합 내 카테고리에 속한다. 예를 들어, CRT, LCD, PDP등 모니터와 관련된 내용을 담고 있는 문서 d 에서 K-NN을 이용한 분류의 결과 후보 카테고리로 c_1 인 컴퓨터 주변기기와 c_{11} 인 모니터가 나왔다고 하자. 즉, $C(d)=\{c_1, c_{11}\}$이고, $C_R(d)=\{c_{11}\}$이다. 따라서, 문서 d 는 c_{11} 인 모니터 카테고리에 속한다.

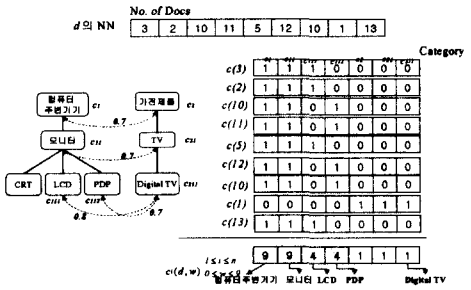
3.2.2. $|C_R(d)|$ 가 2이상이면서 i_1이 동일한 경우

$C_R(d)$ 집합 내 카테고리에 대해 시소러스를 이용하여 문서가 속할 명확한 카테고리를 정한다. 시소러스를 이용하는 방법은 다음과 같다.

1) 문서 d 가 속할 카테고리는 $C_R(d)$ 집합 내 카테고리 중, $C_R(d)$ 에는 속하지 않지만 문서 d 와 관련 있는 카테고리를 고려하여 결정한다.

2) 1)의 방법으로 명확한 카테고리를 정하지 못한 경우, $C_R(d)$ 집합 내 카테고리에 대해 일반화 관계를 가지는 카테고리에 문서가 속한다.

예를 들어, 디지털 TV의 LCD와 PDP 모니터에 관련된 문서 d 가, 시소러스가 <그림 2>와 같이 주어져 있을 때, K-NN을 이용한 분류결과가 다음과 같다고 하자.



<그림 2> $|C_R(d)|$ 가 2이상이면서 i_1이 동일한 경우의 K-NN분류

K가 9, Threshold가 4 일 때, 후보 카테고리로 c_1 인 컴퓨터 주변기기, c_{11} 인 모니터, c_{111} 인 LCD, c_{112} 인 PDP 가 나왔다. 즉, $C(d)=\{c_1, c_{11}, c_{111}, c_{112}\}$ 이고, $C_R(d)=\{c_{111}, c_{112}\}$이다. 그러므로 문서 d 가 $C_R(d)$ 집합 내 c_{111} 인 LCD와 c_{112} 인 PDP 중 어느 카테고리에 속하는지를 명확히 정할 수 없다. 그러므로 $C_R(d)$ 에는 속하지 않지만 $K-NN(d)$ 내의 문서가 포함하는 카테고리인 Digital TV와의 카테고리의 집성화/연관화 관계를 이용하여 LCD 와 PDP 카테고리 중 하나에 문서 d 를 할당한다.

[정의 9] 카테고리 c_i 에 대해 집성화/연관화 관계를 구하는 연산자 $\mathrm{Rel}(c_i)$ 는 다음과 같이 정의한다.

$$\mathrm{Rel}(c_i) = Part(c_i) \cup Ass(c_i)$$

$Part(c_i)$ 는 c_i 와 집성화 관계에 있는 카테고리들의 집합이고, $Ass(c_i)$ 는 c_i 와 연관화 관계에 있는 카테고리들의 집합이다.

[정의 10] c_i 와 $c_j \in \mathrm{Rel}(c_i)$ 간의 관련성을 나타내는 가중치를 $w_{c_ic_j}$ 로 나타내면 다음과 같다.

$$w_{c_ic_j} = Part-of / w_{c_ic_j} + Association-Of / w_{c_ic_j}$$

$Part\text{-}Of / w_{c_ic_j}$ 는 c_i 와 c_j 의 집성화 관계에 대한 관련 정도이고, $Association-Of / w_{c_ic_j}$ 는 c_i 와 c_j 의 연관화 관계에 대한 관련 정도를 나타낸다.

[정의 11] $\mathrm{Rel}(c_i)$ 는 가중치를 고려하여 다음과 같이 재정의된다.

$$\mathrm{Rel}(c_i) = \{c_j / w_{c_ic_j} \mid c_j = Part(c_i) \cup Ass(c_i) \text{ 이고 } w_{c_ic_j} = Part-of / w_{c_ic_j} + Association-Of / w_{c_ic_j}\}$$

예를 들어, <그림 2>에서 $\mathrm{Rel}(Digital\ TV)$ 를 구하면 다음과 같다.

$$Part(Digital\ TV) \cup Ass(Digital\ TV) = \{LCD, PDP\}$$

$$w_{DigitalTV,LCD} = Part-of / w_{DigitalTV,LCD} + Association-Of / w_{DigitalTV,LCD} = 0.8$$

$$w_{DigitalTV,PDP} = Part-of / w_{DigitalTV,PDP} + Association-Of / w_{DigitalTV,PDP} = 0.7$$

$$\mathrm{Rel}(Digital\ TV) = \{LCD/0.8,\ PDP/0.7\}$$

Digital TV 카테고리는 LCD 카테고리와는 0.8 정도의 관련성을 가지고 있고, PDP 카테고리와는 0.7 정도의 관련성을 가지고 있으므로 문서 d 는 c_{111} 인 LCD와 c_{112} 인 PDP의 카테고리 중 LCD 카테고리에 속한다.

만약, 위의 과정을 수행한 후 문서 d가 속할 명확한 카테고리를 정하지 못한 경우는 두 카테고리를 일반화하는 카테고리에 할당한다. $C_R(d)$ 집합 내 카테고리들을 일반화하는 카테고리는 $c_{i_1 i_2 \cdots i_k} \in C_R(d)$ 에 대해 $i_1 \sim i_{k-1}$ 의 값이 동일하고 i_j 의 값이 0인 카테고리이다. 예를 들어, $w_{DigitalTV,PDP}$ 가 $w_{DigitalTV,LCD}$ 와 동일하게 0.8 정도라면, c_{111} 인 LCD 와 c_{112} 인 PDP에 대해 명확한 카테고리를 정할 수 없게 되므로 두 카테고리에 대해 일반화 관계를 가지는 c_{11} 인 모니터 카테고리에 속한다고 결정 하게 된다.

3.2.3 $|C_R(d)|$ 가 2 이상이면서 i_1 이 상이한 경우

$C_R(d)$ 집합 내 카테고리에 대해 집성화/연관화 관계를 가지는 공통적인 카테고리에 문서가 속한다. 예를 들어, 홈시어터와 관련하여 영상기기 및 음향기기에 대한 문서 d 에서 K-NN을 이용한 분류의 결과 후보 카테고리로 c_1 인 영상기기와 c_2 인 음향기기가 나왔다고 하자. 즉, $C(d)=\{c_1, c_2\}$ 이고, $C_R(d)=\{c_1, c_2\}$이다.

따라서, 문서 d 는 c_1 인 영상기기와 c_2 인 음향기기 중 어느 카테고리에 속하는지를 명확히 정할 수 없으므로 두 카테고리에 대해 시소러스의 집성화/연관화 관계를 고려하여 공통적인 카테고리를 구한다. 먼저 영상기기와 음향기기 카테고리에 대해 집성화/연관화 관계를 구하면 다음과 같다.

$$\mathrm{Rel}(\text{영상기기}) = Part(\text{영상기기}) \cup Ass(\text{영상기기}) = \{\text{홈시어터}\}$$

$$w_{\text{영상기기,홈시어터}} = Part-of / w_{\text{영상기기,홈시어터}} + Association-Of / w_{\text{영상기기,홈시어터}} = 0.8$$

$$\mathrm{Rel}(\text{영상기기}) = \{\text{홈시어터}/0.8\}$$

$$\mathrm{Rel}(\text{음향기기}) = Agg(\text{음향기기}) \cup Ass(\text{음향기기}) = \{\text{홈시어터}\}$$

$$w_{\text{음향기기,홈시어터}} = Part-of / w_{\text{음향기기,홈시어터}} + Association-Of / w_{\text{음향기기,홈시어}} = 0.8$$

$$\mathrm{Rel}(\text{음향기기}) = \{\text{홈시어터}/0.8\}$$

영상기기와 음향기기에 대한 집성화/연관화 관계의 공통된 카테고리는 AND연산자를 이용한다.

<영상기기>AND<음향기기>

$$= \mathrm{Rel}(\text{영상기기}) \cap \mathrm{Rel}(\text{음향기기}) = \text{<홈시어터>}$$

따라서, 문서 d 는 후보 카테고리가 아닌 새로운 카테고리인 홈시어터에 속한다.

4. 결론 및 향후 과제

본 논문에서는 통계적인 기법에 전문가의 지식을 바탕으로 한 시소러스를 접목시킴으로써 문서 자동 분류의 성능을 높이는 방법을 제안하였다. 통계적인 분류 기법으로는 분류 알고리즘 중 가장 간단하면서도 80퍼센트 정도의 정확도를 보이고 있는 K-NN 분류 알고리즘을 이용하였고, K-NN을 이용한 문서 분류 시 특정 카테고리로 문서를 분류하기에 애매모호한 경우가 발생하여 명확한 카테고리로 설정을 할 수 없을 때는 전문가에 의해 구축된 객체 지향 시소러스의 집성화/연관화 관계와 일반화 관계를 이용하여 모호성을 줄이는 방법을 제안하였다. 향후 과제로는 제안된 방법을 검증할 실험 절차와 그에 따른 성능 평가가 뒤따라야 할 것이다.

참고문헌

[1] K. Aas and L. Eikvil. "Text Categorization : A Survey" Report NR 941, Norwegian Computing Center, 1999

[2] Wai Lam., Ruiz M., Srinivasan P. "Automatic Text Categorization and Its Application to Text Retrieval" IEEE Transactions on Knowledge & Data Engineering, vol.11, no.6, Nov.-Dec. 1999, pp.865-79. Publisher: IEEE, USA.

[3] Eui-Hong (Sam) Han, George Karypis, and Vipin Kumar. "Text categorization using weight adjusted k-nearest neighbor classification" In Technical Report #99-019, 1999.

[4] 최재훈, 김기헌, 양재동,"객체기반 시소러스 시스템의 설계 및 구현: 반자동화 방식의 구축, 추상화 방식의 개념 브라우징 및 질의기반 참조" 정보과학회 논문지(데이타베이스),Vol. 27, No. 1, pp.64-78, March, 2000.