계층적 분류체계를 위한 자동분류 기법에 관한 연구

An Experimental Study on Text Categorization for Hierarchical Classification

이영숙, 정영미 (연세대학교 문헌정보학과)

Young-Sook Lee, Young-Mee Chung
Dept. of Library and Information Science, Yonsei University

이 연구는 계층적 분류체계를 기반으로 자동분류를 수행할 HiCat 알고리즘을 제안한다. HiCat 알고리즘은 DDC 지식베이스의 주제어와 기계학습을 거친 정보를 동시에 이용하고, 각 계층별로 주제적합성 가중치를 구해 최종 주제범주를 결정한다. 이 알고리즘이 최적의 성능을 보이는 조건을 알아보고, 일반 분류기와의 성능 비교를 통해 HiCat 알고리즘을 평가해 보았다.

1 서론

수작업 분류만큼의 정확성을 획득하면서 시간적인 면에서도 효율성을 가지는 자동분류의 성능을 획득하기 위해서는, 분류 알고리즘과 더불어 문서 분류의 기준이 될 주제 분류체계의 특성도 고려해야 한다.

이 연구에서 제안하는 HiCat(Hierarchical Categorization)은 계층적인 분류체계에 적합한 자동분류 알고리즘이다. HiCat 알고리즘은 DDC 지식베이스와 학습 정보로 구성된 학습 테이블을 동시에 이용하고, 각 계층의 주제범주들과 분류대상 문서와의 주제적합성 가중치를 구하여 최종 주제범주를 결정하는 과정을 거친다. 이러한 과정에서 HiCat 알고리즘이 최적의 성능을 보이는 조건을 알아보고, 기존의 범주화 연구에서 일반적으로 사용되는 분류기와 비교 실험을 통해 성능을 평가해 보았다.

2 계층적 자동분류 실험

2.1 선행연구

Koller와 Sahami(1997)의 연구는 계층적 자동분류와 관련된 여러 논문에서 자주 인용되며 계층구조를 이루는 각 주제범주 단위로 분류작업을 나누는 방법을 제시하였다. 기대 상호 엔트로피로 소수 자질을 선정하였으며, 계층적 자동분류를 위해 자질변수간의 한정된 상호관계를 가진 확장된 나이브 베이즈 분류기를 고안하였다.

Mladenic(1998)는 계층적 자동분류를 위한 주제범주 배경지식으로 야후 디렉토리를 사용하고 나이브 베이즈 분류기를 이용하였다. 계층의 매 주제범주마다 서로 다른 자질리스트를 선정하였고, 특정 계층에 있는 용어를 학습할 때는 그 상위 주제범주에 나타난 이 용어의 학습정보까지 포함시켰다.

DDC 분류체계를 자동분류에 이용한 최근의 연구(Thompson, Shafer, and Vizine-Goetz 1997)는 OCLC의 Scorpion Project (http://orc.rsch.oclc.org:6109)로 DDC 분류체계를 지식베이스로 구축하여 자동분류를 시도하였다. Dolin(1998)과 Larson(1992)은 LCC 분류체계를 이용해 자동분류를 수행하였다. 두 연구 모두 MARC 레코드를 학습집단으로 구성해 각 주제범주를 나타내고 유사도를 이용해 계층

적인 자동분류를 수행하였다.

2.2 실험방법

기본 분류체계는 DDC 분류표를 이용하였으며, DDC 분류번호 300인 사회과학내 39개의 주제범주만을 이용하였다. 실험문서집단은 계층적 분류체계에 따라 문서를 분류하고 있는 인터넷 사이트들을 중심으로 1,110건을 수집하였고, 학습문서 777건, 실험문서 333건으로 분리하였다.

DDC 지식베이스는 DDC 분류표로부터 수집한 여러 정보들을 프레임형식으로 표현하여 구축하였다. 또한 학습테이블은 카이제곱 통계량을 이용한 주제범주별 자질선정 과정을 거친후, 베이즈 확률 공식을 이용해 학습한 정보로 구성하였다.

이렇게 구성된 DDC 지식베이스와 학습테이블은 HiCat 알고리즘을 통해 계층적 자동분류 과정에 이용된다. 계층적 자동분류를 위한 최적 조건을 알아보기 위해 DDC 지식베이스의 탐색방법과 학습테이블의 구축방법을 변수로 실험을 수행하였으며, HiCat 알고리즘 자체도 수작업 분류를 모방한 세가지 방식으로 변형하여 평가해 보았다. 그리고 범주화 연구에서 일반적으로 사용되는 분류기의 성능과 비교 실험을 수행하였다. 성능을 평가하기 위한 평가척도는 정보검색과 범주화 연구에서 주로 사용되는 재현율, 정확률, 정확도, F 척도를 이용하였다.

2.3 HiCat 알고리즘

분류 대상 문서는 단어빈도를 가중치로 한 용어리스트로 표현한다. 추출된 분류 대상 문서의 용어들은 개별적으로 DDC 지식베이스의 주제어, 학습테이블의 자질과 비교 과정을 거쳐, 일치하는 용어가 있으면 그 주제어나 자질이 속한 주제범주 정보와 학습테이블의 확률정보가 주제범주별로 따로 저장된다.

HiCat 알고리즘은 한 문서를 구성하는 개별 용어들이 특정 주제범주를 대표하는 정도를 구하고, 이로부터 분류 대상 문서와 주제범주들 간의 주제적합성 가중치를 계산해 낸다. 모든 주제범주들을 대상으로 주제적합성 가중치를 구한후, 계층별 비교과정을 거쳐 최종 주제범주를 결정한다.

HiCat 알고리즘에서 사용되는 주제적합성 가중치 WR, 지식베이스 가중치 WK, 학습테이블 가중치 WT 의 공식은 다음과 같다.

$$WR_{d_i c_k} = (\alpha \sum_{j=1}^{n} WK_{t_j k} + \beta \sum_{j=1}^{n} WT_{t_j k}) \times \frac{CL}{TL}$$

$$d_i = \{ t_1, t_2, t_3, ..., t_n \}$$

$$WK_{t_j k} = \frac{1}{TC_{t_j}} \times \frac{1}{SP_{c_k}} \times tf_{t_j}$$

$$WT_{t_j k} = P(t_j | c_k) \times 100 \times \frac{1}{TC_{t_j}} \times tf_{t_j}$$

주제적합성 가중치 $WR_{d_i c_k}$는 문서 d_i의 주제범주가 c_k가 될 적합성 정도를 나타낸다. 지식베이스 가중치 $WK_{t_j k}$는 DDC 지식베이스에서 용어 t_j와 일치하는 주제어의 주제범주가 c_k일 경우, 용어 t_j가 주제범주 c_k를 대표하는 정도를 표현한다. 학습테이블 가중치 $WT_{t_j k}$는 학습테이블에서 용어 t_j와 일치하는 학습된 자질의 주제범주 c_k에 대하여 용어 t_j 가 c_k를 대표하는 정도를 나타내는 가중치이다.

주제적합성 가중치 공식의 CL/TL은 계층수준 정보값으로 TL은 계층적 분류체계의 총 계층수, CL은 현재 주제적합성 가중치를 구하고자 하는 계층수준을 나타낸다.

지식베이스 가중치와 학습테이블 가중치 공식의 TC_{t_j}는 지식베이스 및 학습테이블 내에서 용어 t_j가 출현하는 주제범주의 총수이다. SP_{c_k}는 t_j가 출현한 TC_{t_j}개의 주제범주 중에서 c_k와 동일한 상위 노드를 가지는 주제범주의 수이다. tf_{t_j}는 문서 d_i에서 용어 t_j가 출현한 단어빈도로서 분류 대상 문서 d_i에서의 용어 t_j의 중요도를 반영하고 있다. 또한 학습테이블 가중치 공식의 $P(t_j|c_k)$는 용어 t_j가 주제범주 c_k에 속할 조건 확률이다.

주제적합성 가중치식의 α와 β는 지식베이스 가중치와 학습테이블 가중치를 조절할 수 있도록 하는 파라미터이며, 이 연구에서 수행한 실험에서는 두 파라미터의 값이 동일하다.

3 실험결과 및 분석

3.1 DDC 지식베이스 탐색 및 학습테이블 구축방법에 따른 HiCat 알고리즘의 성능

지식베이스의 탐색방법과 학습테이블의 구축방법에 따른 HiCat 알고리즘의 성능변화를 측정하여 최적의 성능을 보이는 조건을 알아보았다. 지식베이스의 탐색방법은 주제어 슬롯의 계층구조에 따른 탐색 유무에 따라 '계층적 탐색', '비계층적 탐색'으로 나뉘어진다. 그리고 학습테이블의 구축방법은 각 주제범주에 속한 학습문서로 학습과정을 거치는 '일반적 구축'과 계층적으로 학습문서를 재구성하여 학습 과정을 거치는 '계층적 구축'으로 나뉘어진다. 이러한 지식베이스의 탐색방법과 학습테이블의 구축방법을 네 가지 경우로 조합하고 각 경우에 HiCat 알고리즘을 적용하였다.

실험 결과, 지식베이스의 계층적 탐색과 학습테이블의 일반적 구축을 동시에 이용한 경우에 가장 나은 성능을 보여주었다. 특히, 지식베이스 탐색방법보다는 학습테이블 구축방법에 따른 성능차가 더 뚜렷하며 원래의 학습문서로 학습테이블을 구축하는 경우가 계층적으로 학습테이블을 구축하는 경우보다 F 척도 10% 이상 우수한 성능을 보였다.

결국, 통제된 정보로 이루어진 지식베이스의 계층적 탐색은 성능향상을 유도하였으나 통제되지 않은 정보들로 구성된 학습테이블의 계층적 구축은 전체 성능은 물론, 특히 상위계층의 성능을 감소시켰다.

3.2 HiCat 알고리즘의 변형

HiCat 알고리즘은 계층별로 각 계층에 속하는 모든 주제범주들을 대상으로 주제적합성 가중치를 구한다. 반면에 계층적인 분류체계에 따른 수작업 분류는 일반적인 주제범주에서 특정적인 주제범주로 단계적인 주제범주 결정방식을 따른다.

이러한 DDC 분류표의 분류번호 체계와 수작업 분류방식을 모방하여 HiCat 알고리즘도 단계적으로 주제범주를 정하도록 세 가지 방식의 변형된 알고리즘으로 실험을 수행하였다. 변형된 분류 알고리즘들은 하향식_1, 하향식_2, 상향식 분류 알고리즘이다.

<표 1> HiCat 및 변형 알고리즘들의 F 척도 비교

	분류 알고리즘	F 척도
기본 알고리즘	HiCat	0.8077
변형 알고리즘	하향식_2	0.5996
	하향식_1	0.5970
	상향식	0.5868

하향식 분류 알고리즘은 계층의 상위부터 주제범주를 결정하여 하위 계층으로 단계적으로 내려가는 방식이고, 상향식 분류 알고리즘은 계층의 최하위부터 상위 계층의 주제범주로 문서를 분류해 가는 방식이다. 하향식_1은 바로 윗 단계에서 결정된 주제범주에 따라 하위 주제범주가 결정되는 반면, 하향식_2는 최상위 계층에서 결정된 주제범주에 속하는 모든 하위 주제범주를 대상으로 분류작업을 수행한다.

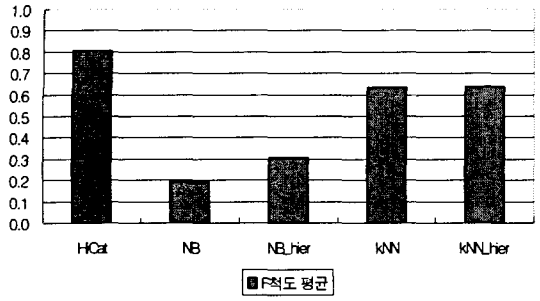
<표 1>은 HiCat 알고리즘과 변형 알고리즘들의 F 척도를 비교한 것이다. HiCat 알고리즘이 전 주제범주를 대상으로 최종 주제범주를 결정하는 방식인 반면, 변형 알고리즘들은 이전 단계의 주제범주 결정이 다음 단계에 영향을 미친다. 따라서 이전 단계의 분류 오류가 최종 주제범주 결정에 영향을 미치게 되어, 세 가지 변형 알고리즘 모두 HiCat 알고리즘보다 낮은 성능을 나타냈다.

3.3 HiCat 알고리즘과 성능 비교용 분류기의 비교

기계학습 분야의 범주화 연구에서 일반적으로 사용되는 나이브 베이즈(naive bayes) 분류기, kNN(k-Nearest Neighbor : kNN) 분류기를 HiCat 알고리즘과 비교, 평가하였다. 또한 나이브 베이즈 분류기와 kNN 분류기를 변형한 NB_hier, kNN_hier의 성능도 함께 평가하였다.

NB_hier는 나이브 베이즈 분류기와 학습과 분류과정은 동일하며, 자질선정 과정만 차이를 두어 주제범주별 상이한 자질리스트를 학습에 이용하였다. kNN_hier는 각 계층별로 k개의 유사한 문서를 골라내어 계층별 주제범주를 결정해 나가는 방식이다.

<그림 1>은 HiCat 알고리즘과 나머지 성능비교용 분류기들의 자질수 변화에 따른 F 척도 평균값을 나타낸 그래프이다. 성능순위는 HiCat 알고리즘이 가장 우수하고 그 다음이 kNN 분류기, kNN_hier, NB_hier순이며 나이브 베이즈 분류기의 성능이 가장 낮았다. 주제범



<그림 1> HiCat 과 성능비교용 분류기의 F척도 평균

주별 상이한 자질리스트로 학습과정을 거치는 NB_hier가 단일 자질리스트로 학습과정을 거치는 나이브 베이즈 분류기보다 F 척도의 성능이 10% 이상 향상되었다. kNN 분류기와 kNN_hier는 거의 동일한 성능을 보여주었다.

전체적으로, HiCat 알고리즘은 나이브 베이즈 분류기나 NB_hier에 비해서 F 척도 61% 이상 나은 성능을 보였으며, kNN 분류기와 kNN_hier 보다도 17% 이상 높은 성능을 나타냈다.

계층적 분류체계의 특성을 고려한 HiCat 알고리즘은 DDC 분류표로부터 추출한 주제어와 학습정보를 동시에 이용함으로써 다양한 주제범주를 대표하는 적절한 용어를 확보할 수 있었다. 또한, 주제범주별 상이한 자질리스트 선정 및 소규모 자질 사용, 주제적합성 가중치를 통해 보다 정확한 분류작업을 할 수 있었다.

4 결론

이 연구에서는 계층적인 분류체계에 적합한 자동분류 알고리즘으로 DDC 지식베이스의 주제어와 학습정보를 동시에 이용하는 HiCat을 제안하였다. 특정 문서와 주제범주들간의 주제적합성 가중치를 이용해 최종 주제범주를 결정하는 HiCat 알고리즘의 성능을 평가하고 최적의 성능 조건을 알아보기 위해 수행한 실험 및 실험결과는 다음과 같다.

첫 번째 실험에서는 지식베이스 탐색 및 학습테이블 구축방법을 변수로 하여, 어떤 경우에 HiCat 알고리즘이 최적의 성능을 나타내는지 알아보았다. 실험결과, 지식베이스의 주제어 슬롯을 계층적으로 탐색하고 각 주제범주에 속하는 학습문서만으로 학습테이블을 구축한 경우가 최적의 성능을 보여주었다.

두 번째 실험에서 HiCat 알고리즘을 계층별 주제결정 방식에 따라 하향식_1, 하향식_2, 상향식으로 변형하여 평가해보았다. 세 변형 알고리즘 모두 기본 HiCat 알고리즘에 비해 성능이 낮게 나타났다.

세 번째 실험에서는 기계학습을 이용한 범주화 연구에서 일반적으로 사용되는 분류기와 HiCat 알고리즘의 성능을 비교하였다. 실험결과 HiCat 알고리즘이 가장 성능이 좋았고, 그 다음이 kNN과 kNN_hier, NB_hier순이었으며, 나이브 베이즈 분류기가 가장 낮은 성능을 보여주었다.

일련의 실험을 통해 계층적 분류체계를 위한 HiCat 알고리즘의 성능이 일반 분류기보다 우수함이 증명되었으며, DDC 지식베이스의 통제된 주제어와 학습정보를 동시에 이용하는 것이 다양한 주제범주로 이루어진 웹문서를 분류하는데 좋은 도구가 됨을 알 수 있었다. 따라서 특정한 분류체계의 구조와 특성을 고려한 분류 알고리즘을 통해 자동분류의 성능을 향상시킬 수 있을 것이다.

참고문헌

Dolin, R.A. 1998. *Pharos: A Scalable Distributed Architecture for Locating Heterogeneous Information Sources.* Ph.D. diss., University of California Santa Barbara.

Koller, D., and Mehran Sahami. 1997. "Hierarchically Classifying Ducuments Using Very Few Words". ICML97: 170-178.

Larson, R.R. 1992. "Experiments in Automatic Library of Congress Classification". JASIS,43(2): 130-148.

Mladenic, D. 1998. *Machine Learning on non-homogeneous, distributed text data.* Ph.D. diss., University of Ljubljana.

Thompson, R., Keith Shafer, and Diane Vizine-Goetz. 1997. "Evaluating Dewey Concepts as Knowledge Base for Automatic Subject Assignment".