협력적 필터링을 위해 연관 단어 빈도를 이용한 웹 문서 분류

하원식*○, 정경용*, 정헌만*, 류중경*, 이정현**
인하대학교 컴퓨터 · 정보공학과*, 인하대학교 컴퓨터공학부**
{sigboy, dragon}@nlsun.inha.ac.kr, hmjung@inhatc.ac.kr, jkryu@daelim.ac.kr, jhlee@inha.ac.kr

Classification of Web Documents Using Associative Word Frequency for Collaborative Filtering

Won-Sik Ha*○, Kyung-Yong Jung*, Heon-Man Jung*, Joong-Kyung Ryu*, Jung-Hyun Lee**
Dept. of Computer Science & Information Engineering, Inha University*
School of Computer Science & Engineering, Inha University**

요 약

기존의 웹 문서 분류 시스템에서는 많은 시간과 노력을 요구하며, 연관 단어가 아닌 단일 단어만으로 웹 문서들을 분류하여 단어의 중의성을 반영하지 못해 많은 오분류가 있었다. 이러한 문제점을 해결하기 위해 본 논문에서는 협력적 필터링을 위한 연관 단어 빈도를 사용한 웹 문서 분류 방법을 제안한다. 제안된 방법에서는 웹 문서 내에서 단어들을 추출하고 빈도 가중치를 계산한다. 추출된 단어를 Apriori 알고리즘에 의해 연관 규칙을 생성하고 신뢰도에 단어 빈도 가중치를 반영한다. 수정된 신뢰도를 ARHP 알고리즘에 적용하여 연관 단어들 사이의 유사정도를 계산하고 유사 클래스를 구성한다. 생성된 유사 클래스들을 기반으로 웹 문서를 α-cut을 이용하여 분류한다. 성능평가를 위해 기존의 문서 분류 방법들과 비교 평가를 하였다.

1. 서 론

기존의 문서 분류 방법에는 다변량 회기모델, 최근인접분류, 베이지안 확률접근, 의사결정트리, 신경망, 기호규칙학습, 연역학습 알고리즘 등의 많은 웹 문서 분류 방법들이 제안되었다. 이러한 방법들은 특징 공간이 다차원으로 표현되고, 가장 성능이 뛰어나다고 알려져 있는 베이지안 확률 모델[1] 같은 경우 특징에 대한 독립가정을 하고 있어 웹 문서 분류의 정확도를 저하시키지 않으면서 웹 문서를 분류하기가 쉽지 않다. 본 논문에서는 이러한 문제점을 보완하기 위해 협력적 필터링을 위한 연관 단어 빈도를 사용한 웹 문서 분류 방법을 제안한다. 기존의 웹 문서 분류 방법들이 가지고 있는 중의성 문제를 해결하기 위해 문서의 특징을 연관 단어 형태로 추출하고, 단어의 발생 빈도에 따른 가중치를 반영하여 웹 문서 분류 방법의 성능을 개선하였다. 제안된 방법의 성능평가를 위해 기존의 문서 분류 방법들과 비교 평가하였다.

2. 관련 연구

2.1 정보이득

정보이득은 기계학습의 분야에 있어서 단어 선정의 기준으로써 자주 사용된다. 문서 안에 있는 단어의 출현과 비출현을 알아냄에 의해 카테고리 예측을 위해 획득한 정보의 비트의 수를 측정한다. $\{c_1, c_2, \ldots c_m\}$는 목적 공간에서의 카테고리 집합이라고 할 경우 단어 t의 정보이득은 식(1)와 같다.

$$G(t) = \sum_{i=1}^{m} \Pr(c_i)\log\Pr(c_i) + \Pr(t)\sum_{i=1}^{m}\Pr(c_i \mid t)\log\Pr(c_i \mid t) + \Pr(\bar{t}) \sum_{i=1}^{m}\Pr(c_i \mid \bar{t})\log\Pr(c_i \mid \bar{t}) \quad (1)$$

2.2 상호정보

상호정보는 통계적인 언어에서 단어간의 연관 관계를 설정하는데 사용된다. 단어 t와 카테고리 c의 관계를 고려할 경우 A는 단어t와 c가 공기하는 횟수이며, B는 c가 나타나지 않고 t만이 나타난 수이며, C는 t없이 c만 나타난 수이며, N은 문서의 전체수이다. 이러한 경우 t와 c사이의 상호정보는 식(2)과 같다.

$$I(t,c) \approx \log \frac{A \times N}{(A+C)\times(A+B)} \quad (2)$$

2.3 용어 연관도

용어 연관도는 문서 검색에서 어휘 감소를 위해 제안된 방법이다. 이 방법은 긴밀하게 관련된 문서에서 한 단어가 얼마나 자주 나타나는가에 기반을 두고 단어를 평가한다. 이것은 두 문서의 유사도가 임계값 이상이라는 가정에서 훈련 집합에서 측정한다. 용어 연관도는 하나의 단어가 주어진 관련된 문서의 쌍 중에서 각각의 쌍에 나타날 조건 확률을 기반으로 측정된다. 단어 t에 대한 용어 연관도는 식(3)과 같다. 식(3)에서 x와 y는 관련된 문서를 나타낸다.

$$s(t) = \Pr(t \in y \mid t \in x) \quad (3)$$

2.4 Naive Bayes 분류자

Naive Bayes 분류자는 문서에 나타나는 모든 단어를 특징으로 추출한다. Naive Bayes 분류자는 실험문서 D의 특징이 $\{n_1, n_2, \ldots\ldots n_m\}$라고 하였을 때 식 (4)에 의해서 $\{class1, class2, \ldots\ldots, classN\}$중 하나의 클래스로 분류한다.

$$class = \arg^{N}_{ID=1} \max\ P(classID)\prod_{k=1}^{m} P(n_k \mid classID) \quad (4)$$

대용량 데이터베이스에서는 성능이 뛰어난 편이지만, 각 단어는 문맥에 관계없이 독립임을 가정하고 있어서 단어의 중의성 문제를 가지고 있으며, 계산과정이 매우 복잡하여 시간자원의 손해를 감수해야 하는 단점이 있다.

3. 연관 단어 빈도를 사용한 웹 문서 분류 방법

본 논문에서 제안하는 연관 웹 문서의 분류 방법은 그림 1과 같다. EachMovie 데이터 셋[2]에 포함된 URL의 웹 페이지에서 불용어들을 제거하고 WebBot[3]을 이용하여 단어들을 추출한다. 추출된 단어들을 기반으로 각 단어들의 빈도 가중치를 계산하고, Apriori 알고리즘을 이용하여 연관 규칙을 생성한다. 생성된 연관 규칙의 신뢰도에

빈도 가중치를 반영하여, 연관 단어들간에 유사클래스를 생성하기 위한 ARHP 알고리즘의 가중치로 사용한다. 생성된 유사 클래스들을 기반으로 웹 문서들을 α-cut을 이용하여 분류한다.

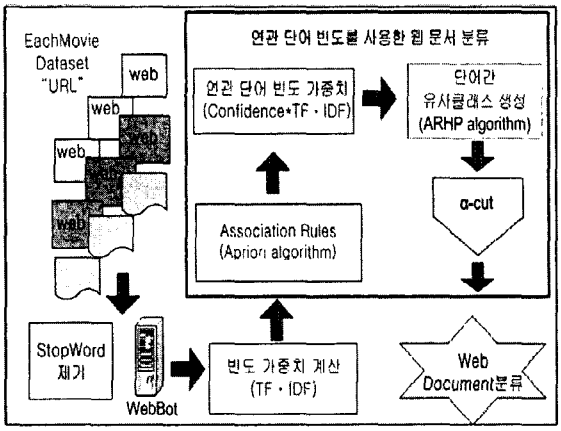


그림 1. 제안된 웹 문서 분류 방법

3.1 연관 단어들간의 빈도 가중치

임의의 웹 문서에서 특정 단어의 발생빈도가 높으면 높을수록 그 단어는 중요한 의미를 가지고 웹 문서의 주제를 많이 반영하고 있다고 할 수 있다. 본 논문에서는 단어의 중의성 문제를 해결하기 위해 연관 단어들간의 유사도를 이용하여 웹 문서를 분류한다. 단순히 단어의 연관성만으로 웹 문서를 분류한다면, 그 단어의 발생빈도는 반영하지 못하게 된다. 따라서 정보검색 분야의 TF·IDF 기법을 사용하여 단어의 빈도가중치를 반영한다. 일반적인 TF·IDF 기법은 식(5)와 같이 나타낼 수 있다[4].

$$W_{ik} = f_{ik} \times [\log(n) - \log(DF) + 1] \qquad (5)$$

본 논문에서는 연관 단어들간의 빈도 가중치를 반영하기 위해 식(5)를 이용한다. 식(6)은 연관 단어간의 빈도 가중치로서 각 연관 단어의 TF·IDF 값의 평균을 사용한다.

$$W' = \frac{1}{n}\{\sum^{n}(f_{ik} \times [\log(n) - \log(WF) + 1])\} \qquad (6)$$

위 식(6)에서 n은 단어의 수, f_{ik}는 단어의 빈도수, WF는 단어가 나타난 웹 문서의 수이다.

3.2 연관규칙 생성

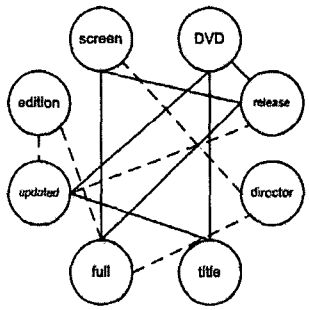
Apriori 알고리즘은 지지도를 이용하여 동시에 자주 발생하는 아이템(frequent item)들을 정제하고, 빈발 아이템 집합에서 생성된 규칙들은 신뢰도를 이용하여 정제하는 방식이다. Apriori 알고리즘은 후보 아이템 집합에서 각각의 지지도를 계산한 후 사용자가 정의한 지지도보다 크거나 같은 조건을 만족하는 데이터로 빈발 아이템 집합(Large itemset)을 구성한다. 그리고 후보 아이템 집합들은 전 단계의 빈발 아이템 집합의 조인 연산을 통해 구성된다. 식(7)의 지지도(Support)는 연관 규칙을 반영하는 트랜잭션이 전체 데이터베이스에서 얼마만큼의 비율을 차지하고 있는지를 나타내는 측정 기준으로서 통계적 중요성을 반영한 것이다. 식(8)의 신뢰도(Confidence)는 규칙이 실제로 정확한지를 판단하는 정도로서, 연관 단어들 사이의 강도를 나타내는 측정 기준(결합도)으로 사용한다. 본 연구에서는 중의성 문제를 해결하기 위하여 신뢰도만을 사용한다.

$$\text{Support} = \frac{\#\,Tuple\ containing\ both\ A\ and\ B}{total\ \#\ of\ tuples} \qquad (7)$$

$$\text{Confidence} = \frac{\#\,Tuple\ containing\ both\ A\ and\ B}{\#\,Tuple\ containing\ A} \qquad (8)$$

3.3 연관 단어들간 유사도 계산

연관 규칙을 기반으로 하여 ARHP(Association Rule Hypergraph Partitioning) 알고리즘으로 연관 단어들간의 유사도를 계산한다. ARHP 알고리즘은 연관규칙과 Hypergraph Partitioning을 이용하여 트랜잭션 기반의 데이터베이스에서 연관된 항목들을 클러스터링하는 방법이다[5]. Hypergraph Partitioning을 위한 가중치로는 식(8)의 연관규칙의 신뢰도에 식(6)의 연관 단어들간의 빈도 가중치를 적용한다. 그림 2는 연관 단어들간의 유사도를 계산하는 과정이다.



생성된 유사 클래스

{screen,full,release}

{DVD,updated,title}

그림 2. 연관 단어들간 유사도 계산

Hypergraph H={V,E}는 단어들로 구성된 정점(vertex)들의 집합 V와 빈번한 항목집합들을 나타내는 하이퍼 간선들의 집합 E로 구성된다. 즉, 연관 규칙에 포함되는 항목들을 정점으로, 연관 관계를 하이퍼 간선으로 매핑하는 것이다. 연관 단어들 간의 유사도를 이용하여 Hypergraph Partitioning으로 유사 클래스를 생성한다.

3.4 연관 웹 문서 분류(α-cut)

α-cut은 소속함수의 [0,1]사이의 값에서 임의의 α ($0 \le \alpha \le 1$)값이 되는 함수 값에 대한 퍼지의 상태 변수의 구간을 나타낸다. 이 α-cut은 퍼지 집합의 원소들에 대해 집합에 속할 기준을 정의할 때 사용된다[6]. 연관 단어들을 원소로 하는 연관 문서들에 대해서 임의의 [0,1]의 값을 가진 α-cut을 적용한 웹 문서 집합 $WDoc_\alpha$ 는 식(9)과 같이 나타낼 수 있다.

$$WDoc_\alpha = \{x | WDoc(x) \ge \alpha\} \quad (x\text{는 연관단어}) \qquad (9)$$

따라서 웹 문서 집합 $WDoc_\alpha$ 는 유사 클래스에 속할 소속정도 값이 α 값 이상으로 이루어진 집합이다. 본 논문에서 제안하는 방법은 웹 문서 집합에서 추출한 단어들간의 연관 규칙을 적용하여 웹 문서를 분류함으로써, 사용자의 관심도를 보다 많이 표현하고, 키워드 단순 매칭에 의한 검색기법의 단점인 의미상으로 연결된 문서에 대한 분류를 해결한다. 웹 문서 간의 유사정도에 따라 문서를 분류하는 방법은 웹 문서에서 생성된 단어의 유사 클래스를 이용하여 동의어 문제를 해결할 수 있도록 웹 문서를 분류한다. 따라서 제안한 방법은 사용자의 관심을 보다 효율적으로 반영하고 의미적으로 관련 있는 웹 문서를 동일한 카테고리로 분류함에 따라 더 정확한 문서 분류를 할 수 있다.

3.5 분류된 웹 문서 기반의 협력적 필터링

아이템 기반의 협력적 필터링은 서로 다른 두 아이템에 대해 동시에 평가한 사용자들의 평가치를 기반으로 아이템간의 유사도를 계산하여 추천하는 방법이다. 아이템 기반의 협력적 필터링의 군집 단계에서 제안한 방법으로 분류된 웹 문서를 이용한다. 각 영화를 대표하는 웹 문서들을 아이템으로 간주하고, 사용자들의 영화에 대한 평가치로 유사도를 계산하여 추천한다. 웹 문서 분류시 중의성을 해결하였고, 각 영화의 주제를 잘 반영하고 있는 자주 등장하는 단어들을 기반으로 하여 분류된 웹 문서로 추천을 함으로써 기존의 시스템들에 비해 보다 정확한 결과를 얻을 수 있다.

4. 실험 및 성능 평가

본 논문에서 제안한 연관 단어 빈도를 사용한 웹 문서 분류 방법은 Visual C++ 6.0으로 구현되었으며, 실제 실험 환경은 Pentium4 1.5Ghz, 256MB Ram 환경에서 수행되었다. 컴팩 연구소에서 18개월 동안 협력적 필터링 알고리즘을 연구하기 위해서 영화에 대한 사용자의 선호도를 조사한 EachMovie 데이터 셋[2]을 사용한다. 이 데이터는 총 72,916명의 사용자와 1,628종류의 영화에 대해서 0.0에서부터 1.0까지 0.2 간격으로 명시적으로 평가한 선호도로 구성되어 있다. 영화의 특징은 10개의 장르로 구분되어 있고, 각 영화를 대표하는 URL이 함께 제공된다. 실험 방법은 3가지로 구분하였다. 첫번째 방법(DF+NB)은 기존의 TF·IDF 기술에 Naïve Bayes 분류자를 적용한 방법이고, 두번째 방법(IG+NB)은 정보이득 기술로 문서의 특징을 추출하여 Naïve Bayes 분류자로 문서를 분류하였다. 마지막 방법(AWF)은 본 연구에서 제안한 연관 단어 빈도를 사용한 웹 문서 분류 방법을 적용하였다.

4.1 실험 방법 및 결과

실험을 위하여 EachMovie 데이터 셋을 전처리 하여 30,861명의 사용자와 1,612 종류의 영화에 대해서 실험을 진행하였다. EachMovie 데이터 셋에 있는 각 영화를 대표하는 URL을 대상으로 WebBot을 이용하여 단어를 추출한다. 추출한 단어들을 대상으로 TF·IDF 값을 계산한다. 표1은 각 단어별 TF·IDF추정치 값을 나타낸다.

표 1. 각 단어별 TF·IDF 추정치

웹문서 / 단어	$WDoc_1$	$WDoc_2$	$WDoc_3$	$WDoc_4$	$WDoc_5$	$WDoc_6$	...
guide	0.527	0.128	0.419	0.721	0.564	0.183	...
set	0.397	0.093	0.108	0.518	0.927	0.429	...
demand	0.426	0.286	0.527	0.259	0.714	0.112	...
network	0.776	0.583	0.288	0.158	0.627	0.576	...
space	0.115	0.719	0.624	0.715	0.253	0.426	...
...							...

추출된 단어를 Apriori 알고리즘을 이용하여 연관 단어로 구성한다. 생성된 연관 단어들의 신뢰도에 각 단어별 빈도를 반영한다. 연관 단어의 중복되는 신뢰도는 평균값을 사용하고[5], 식(6)에 의한 연관 단어 빈도를 신뢰도에 반영한다. 표 2는 수정된 신뢰도 값을 나타낸다.

표 2. 연관 단어 빈도를 반영한 신뢰도 값

추출된 연관 단어	수정된 신뢰도
[title][DVD][search][match][choosing][best]...	81.76%
[nothing][city][another][miss][night][woman]...	47.52%
[change][more][diligent][awareness][like][ape]...	76.29%
[strong][sound][beauty][zoo][somebody][clerk]...	69.52%
[record][all][learn][orange][score][total][boom]...	84.49%
...	...

표 2의 수정된 신뢰도 값을 ARHP 알고리즘의 가중치로 사용하여 연관 단어들간 유사도를 계산하고 유사 클래스를 생성한다. 표3은 생성된 유사클래스를 나타낸다.

표 3. 생성된 유사 클래스 리스트

클래스 번호	클래스에 포함된 연관단어
Class 1	[moon][night][virgin][weird][speed] [exit][child]...
Class 2	[city][another][bell][drink][mall][coffee][building]...
Class 3	[user][cell][score][extreme][winter][admire][tip]...
Class 4	[air][time][retire][die][useful][equipment][narrow]...
...	...

유사 클래스 내에 속한 단어들을 가진 웹 문서들은 표 4와 같이 α-cut을 이용하여 분류된다. 본 연구에서는 실험을 통하여 가장 좋은 성능을 보인 0.6-cut을 사용한다.

표 4. 유사 클래스에 속한 웹 문서들 분류(클래스 번호가 1인경우)

웹문서 / 단어	$WDoc_1$	$WDoc_2$	$WDoc_3$	$WDoc_4$	$WDoc_5$	$WDoc_6$	...
moon	0.513	0.283	0.591	0.716	0.737	0.438	...
night	0.327	0.429	0.743	0.246	0.671	0.611	...
virgin	0.221	0.294	0.418	0.167	0.917	0.794	...
weird	0.015	0.213	0.726	0.491	0.478	0.616	...
speed	0.045	0.671	0.612	0.624	0.691	0.815	...
exit	0.167	0.543	0.595	0.038	0.597	0.577	
...							...

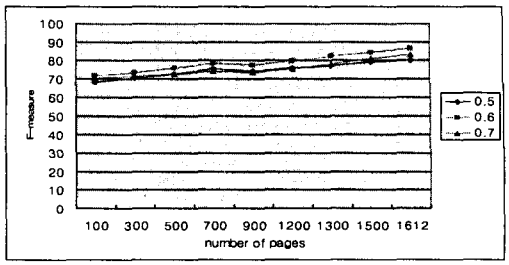


그림 3. α값에 따른 성능비교

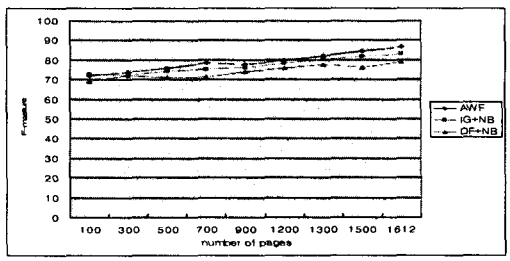


그림 4. 기존 방식과의 성능비교

그림 3은 α값에 따른 성능비교 그래프이고, 그림 4는 기존 방식과의 성능비교 그래프이다. 분류의 측정은 정확도와 재현율을 이용한 F-measure 측정식[4]으로 성능을 평가한다. 그림 4를 보면 기존의 방식과 비교해 볼 때 분류의 성능이 뛰어나다는 것을 알 수 있다. 연관 단어 빈도를 사용한 문서 분류의 성능은 86.48%로 정보이득을 사용한 방법보다는 3.45%, TF·IDF를 사용한 방법보다는 7.33% 높다. 전체적으로 연관 단어 빈도를 사용한 웹 문서 분류 방법이 기존의 다른 방법보다 성능이 우수함을 알 수 있다.

5. 결론

본 논문에서는 EachMovie 데이터 셋에 포함된 URL의 웹 페이지에서 단어들을 추출하고, 빈도 가중치를 계산하여 생성된 연관규칙의 신뢰도에 반영하고, 유사클래스를 생성시켜 웹 문서들을 α-cut을 이용하여 분류하는 방법을 제안하였다. 제안한 방법에 대한 성능을 기존의 분류 방법들과 비교 실험한 결과 분류의 정확도가 향상되었기 때문에 본 논문에서 제안한 방법이 효과적임을 알 수 있었다.

향후 과제로는 분류된 웹 문서들을 아이템 기반의 협력적 필터링 기술에 적용할 수 있다면 보다 더 효율적인 추천 시스템을 기대할 수 있을 것이다.

Acknowledgement

본 연구는 정보통신부 대학 IT 연구센터 육성 지원사업의 연구결과로 수행되었습니다.

참고문헌

[1] T. Michael, *Machine Learning*, McGraw-Hill, pp.154-200, 1997.

[2] P. McJones, EachMovie collaborative filtering dataset, URL:http://www.research.digital.com/SRC/eachmovie, 1997.

[3] 인하대학교, 사용자 중심의 지능형 정보검색 시스템, 최종 연구 개발 보고서, 정보통신부, 1997.

[4] R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, pp.29-83, 1999.

[5] E. Han, G Karypis, V. Kumar, B. Mobasher, "Clustering Based On Association Rule Hypergraphs," In Proc. SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.

[6] L. T. Koczy, *Information retrieval by fuzzy relations and hierarchical co-occurrence*, 1997.