구조 정보를 이용한 웹 문서 범주화 모형

A Categorization Model Based On Information Structure of HTML Documents

조이영, 최상희, 정영미 (연세대학교 문헌정보학과)

Yi-Young Cho, Sang-Hee Choi, Young-Mee Chung
Department of Library and Information Science, Yonsei University

본 연구는 다양한 웹 문서를 효과적으로 범주화 할 수 있는 모형을 구축하는데 그 목적이 있다. 이를 위해 본 연구에서는 웹 문서가 가지고 있는 구조 정보인 링크(link)와 문서 단계(level)를 활용하여 문서 유형을 식별한 후, 각 유형별로 범주화 과정을 달리 적용하여 범주화 성능을 개선시키는 방법을 고안하였다.

1. 서론

웹 문서 범주화는 웹 문서에 미리 정해진 범주를 자동으로 부여하는 것이다. 웹 문서 범주화는 문서를 대상으로 한다는 점에서 기존의 문서 범주화와 동일한 맥락을 가진다. 그러나 비교적 평면적인 구성을 가진 일반 문서에 비해 웹 문서는 링크라는 문서간 계층적인 연결 구조를 가지면서 정보를 단계적, 다차원적으로 제공하고 있다는 점에서 많은 차이가 있다. 링크와 문서 태그(tag) 등 다양한 요소로 구성된 웹 문서는 연결 구조와 보유 정보의 특성에 따라 몇 가지 유형으로 분류될 수 있다(Pirolli 1996). Lewis(1992)는 이와 같은 웹 문서의 특성을 분석하여 웹 문서 하나에 여러 가지 이질적인 정보가 혼재되어 있는 경우 연결 정보를 이용, 내용을 분할함으로써 문서당 할당될 범주의 수를 선정하는 접근방법을 제시하였다.

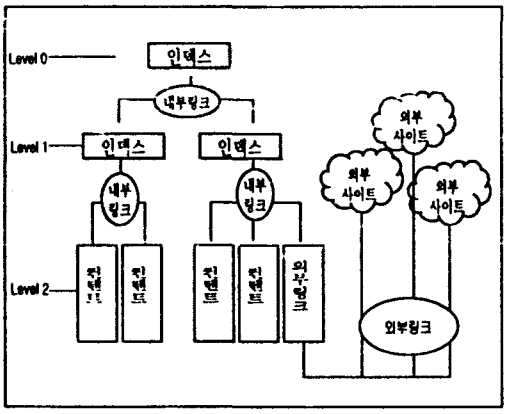
본 연구에서는 이와 반대되는 접근방식으로 문서가 범주화에 유용한 정보를 충분히 가지고 있지 못한 유형일 경우 연결 정보를 이용하여 내용을 보충함으로써 범주화의 정확률을 개선시키고자 하였다. 범주화를 하기 이전에 웹 문서를 유형별로 식별하여 범주화 성능이 낮게 나타나는 문서 유형은 전처리를, 한 후 범주화를 하고, 범주화 성능이 높게 나타나는 문서 유형은 그대로 범주화에 적용하는 방법을 고안하였다.

2. 웹 문서의 특성 및 유형

다양한 환경에서 정보 교환이 가능하도록 하기 위해, 웹 문서에는 전달하고자 하는 내용에 덧붙여 문서의 구조적 성질 및 연관된 다른 문서와의 상호 관계도 기술된다. 웹 문서 작성에 사용되는 기술(記述)언어 HTML(HyperText Markup Language)은 정보의 위치, 형태 및 다른 문서와의 연결관계를 구조적으로 표현한다.

한 사이트 내에서 웹 문서는 링크(link)라는 계층적인 연결 구조를 가진다. 링크는 보통 동일 사이트 내부의 문서를 연결하는 내부링크와 외부 사이트의 문서로 연결되는 외부링크로 나눌 수 있다. 내부링크의 연결 구조를 계층적으로 알려주는 것이 단계(level)이다. 예를 들자면 한 사이트의 홈페이지는 level 0에 해당하고 홈페이지에서 연결되는 메뉴들은 level 1에 해당한다고 할 수 있다. 이 연결 구조에 따른 웹 문서 유형을 구체화하면 아래와 같다(그림 1 참조).

- 인덱스(index) : 관련 정보를 가지고 있는 사이트 내부의 여러 문서로 연결되도록 하는 문서
- 컨텐트(content) : 실제 정보를 가지고 있는 문서
- 외부링크 모음(external links) : 관련 있는 다른 외부 문서로 안내하는 문서, 예를 들면 추천 사이트나 관련 링크 모음 등이 이에 해당한다.



< 그림 1 > 한 웹사이트 내에서 연결 구조로 본 웹 문서 유형과 단계

일반적으로 상위 단계로 올라갈수록 인덱스 문서가 많고 하위 단계로 갈수록 컨텐트 문서가 많다. 본 연구 결과 level 4 이하의 문서는 대부분 컨텐트 문서인 것으로 밝혀졌다.

3. 웹 문서 2단계 범주화 모형 실험

3.1 실험대상 및 방법

본 실험 대상인 웹 문서 450건은 디렉토리 서비스인 "네이버"의 기업 디렉토리 중 '컴퓨터' 아래 분류된 웹사이트에서 수집되었다. 수집된 웹 문서는 대부분 보안, 컨설팅, 네트워크, 시스템 통합 분야의 문서였다. 이 분야 기업사이트에 속한 웹 문서는 회사소개, 보도자료, 제품안내, 관련 기술 정보, 공지사항에 대한 것이 많았다. 전체 제공 정보량에서 각 사이트간 다소 차이가 있었다. 그러나 정보를 평균치 이상으로 많이 제공하는 사이트들은 제공 정보의 다수가 다른 외부 사이트의 문서였으므로 대부분 외부링크에 해당되었다. 그러므로 사이트내에서 자체제공하는 정보량에서는 큰 차이가 없었다.

수집된 웹 문서는 세 가지 측면에서 활용되었다. 첫째, 자동으로 웹 문서 유형을 식별하는 모듈에 적용할 수 있는 규칙을 도출하기 위하여 구조적으로 분석하는데 이용되었다. 둘째, 본 실험에서는 문서 벡터(vector)와 주제 벡터 간의 유사도를 측정하여 범주화하고자 하였으므로 범주화 기준이 되는 주제 벡터 생성에 활용되었다.

주제 벡터는 학습집단 웹 문서 250건을 자동으로 클러스터링한 후 수작업으로 재조정하여 생성하였다. 최종 선정된 주제 범주의 센트로이드 벡터(centroid vector)는 모두 21개이다. 마지막으로 실제 범주화하여 정확률을 평가하는 실험집단으로 200건의 웹 문서가 이용되었다. 이번 실험에서는 실험대상을 컨텐트와 인덱스로 제한하였다. 외부링크 모음 유형으로 식별된 문서는 인덱스나 컨텐트 같은 문서 유형과는 다른 특성을 가지고 있으므로, 범주화에서도 컨텐트와 인덱스 처리 방식과 다른 접근 방식을 택하는 것이 적절할 것이다(조광제 1997).

웹 문서 범주화는 코사인 유사계수를 적용한 벡터간 유사도(vector similarity) 측정 방법을 사용하였다. 이 방법은 범주화하려는 문서와 주제 범주를 각각 자질(색인어) 벡터로 구성하고, 두 벡터 사이의 유사도를 비교하여 유사도가 가장 높은 범주를 문서에 할당하는 방법이다. 예를 들어 문서 D가 T_1, T_2, T_3, T_6의 자질을 가지고 있고, 주제 범주 C_j가 T_1, T_2, T_6, T_7, 의 자질을 가지고 있다면 문서 D와 범주 C_j의 벡터는 D=(1,1,1,0,0,1,0), C_j=(1,1,0,0,0,1,1)이 된다. 두 벡터간의 유사도는 다음과 같은 공식을 사용하여 계산된다.

$$S(D_i, C_j) = \frac{\sum_i D_i \sum_j C_j}{\sqrt{\sum_i D_i^2 \sum_j C_j^2}}$$

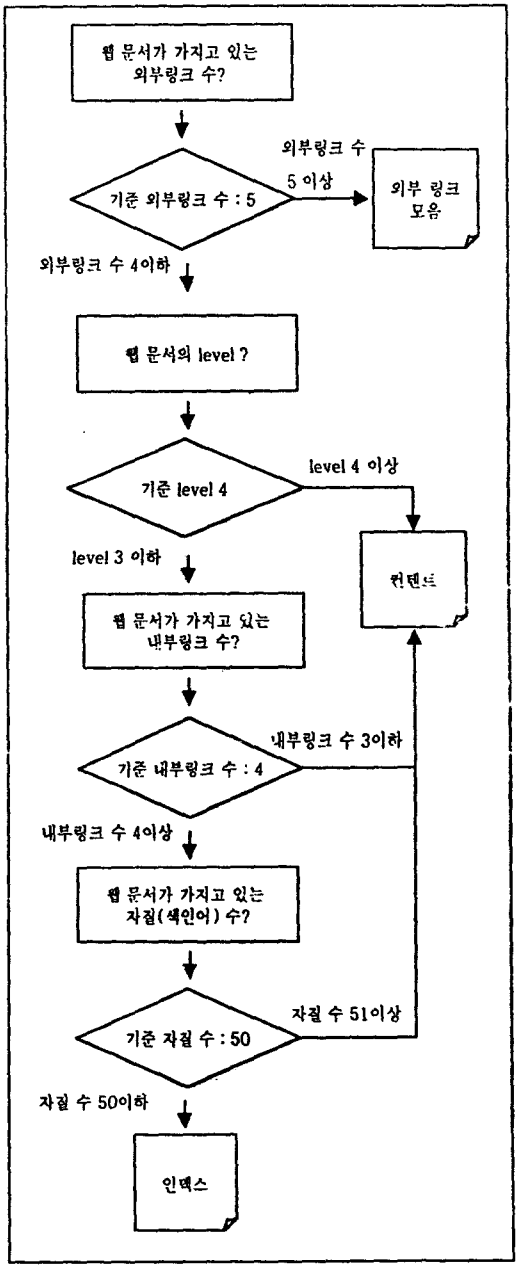
1) 웹 문서 유형 식별

본 연구에서 규정한 웹 문서의 유형은 앞에서 설명하였듯이 컨텐트, 인덱스, 외부링크 모음이다. "네이버"의 기업 디렉토리 중 '컴퓨터' 아래 분류된 웹사이트에서 링크와 문서 단계를 분석하여 도출된 웹 문서 유형 식별 알고리즘과 기준은 그림 2, 표 1과 같다.

식별조건 / 문서유형	내부 링크수	외부 링크수	문서단계	자질수
인덱스	4 이상	4 이하	0~3	50 이하
컨텐트	3 이하	4 이하	0~3	51 이상
			4 이상	
외부링크 모음		5 이상		

< 표 1 > 웹 문서 유형 식별 기준

도출된 기준을 본 실험 대상 웹 문서 중 389건에 적용하여 자동 식별한 결과, 「컨텐트」 문서는 91%의 식별 정확률을, 「외부링크 모음」 문서는 96%의 식별 정확률을 보였다. 이 두 유형의 문서가 90%가 넘는 식별 정확률을 나타낸 반면, 「인덱스」 문서는 81%의 가장 낮은 식별 정확률을 보였다. 문제가 되는 웹 문서 유형인 인덱스 문서는 규칙성을 도출하기가 쉽지 않았다. 그 이유는 인덱스가 가장 다양한 형태로 나타나는 문서 유형이었기 때문이다.



< 그림 2 > 웹 문서 유형 식별 과정

2) 문서 확장

컨텐트와 인덱스의 특성을 조사하는 과정에서 사전에 유형별로 문서를 분류하여 자동 클러스터링한 실험 결과를 분석, 응용하였다.

사전 실험은 국내 대학 컴퓨터공학과 웹사이트에서 수집한 웹 문서 113건을 대상으로 수행되었는데, 문서 유형이 클러스터링 정확률에 영향을 미치는지 조사한 것이다. 실험집단은 인덱스 문서 43건, 컨텐트 문서 70건으로 구성되었다. 클러스터링 결과를 살펴보면, 인덱스 문서의 수가 컨텐트 문서보다 적음에도 불구하고 형성된 클러스터의 수는 컨텐트 문서의 클러스터 수 보다 더 많았다. 또한 클러스터링 정확률도 컨텐트 문서보다 9% 정도 떨어지는 양상을 보였다(표 2 참조).

문서유형	문서수	형성된 클러스터수	정확률
인덱스	43	11	81%
컨텐트	70	10	90%

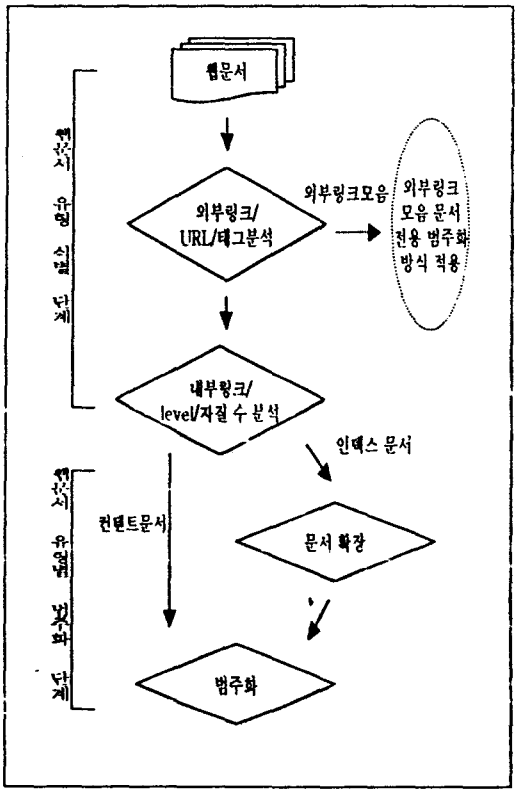
< 표 2 > 웹 문서 클러스터링 실험결과

이와 같은 양상이 나타난 이유 중 하나는 인덱스 문서가 궁극적으로 제공하고자 하는 정보를 충분히 표현하고 있지 못하기 때문이라고 할 수 있다.

따라서 본 실험에서는 인덱스 문서가 지시하는 컨텐트 문서로 인덱스 문서를 확장하여 범주화하였을 때, 인덱스 문서만 가지고 범주화한 결과보다 더 좋은 분류 정확률을 획득할 수 있는지 알아보고자 하였다. 인덱스 문서를 확장하는데 사용한 핵심요소는 웹 문서가 가지고 있는 특성 중 하나인 구조정보로서 내부링크이다. 인덱스 문서 본문에 있는 태그를 이용하여 링크를 인식하고, URL을 비교하여 내부링크를 식별한 후 내부 링크로 연결된 문서로 확장하였다. 이와 같이 웹 문서의 전처리가 범주화에 미치는 영향을 고찰하기 위하여, 컨텐트와 인덱스로 식별된 실험집단 웹 문서 200건을 대상으로 컨텐트 문서는 그대로 범주화하고 인덱스 문서는 문서 확장을 한 경우와 하지 않은 경우로 구분하여 범주화 한 다음 결과를 분석, 범주화 성능을 측정하였다.

3) 웹 문서 범주화 실험

본 실험에서는 웹 문서를 2단계에 걸쳐 처리하는 접근 방식을 시도하였다(그림 3 참조).



< 그림 3 > 웹 문서 2단계 범주화 모형

첫 단계는 실험할 웹 문서 유형인 컨텐트와 인덱스를 선별하고 실험대상이 아닌 외부링크모음을 분리해내는 「문서 유형 식별 단계」이다. 이 단계에서 세 가지 유형의 웹 문서 식별 정확률은 89%였다. 유형이 식별된 문서는 유형별로 다음 과정인 「범주화 단계」로 넘어가게 된다. 컨텐트는 이미 범주화에 유용하게 적용

될 수 있는 정보를 충분히 보유하고 있는 것으로 가정하여 그대로 범주화하였다. 그러나 인덱스는 궁극적으로 제공하고자 하는 컨텐트 문서로 내용을 확장하는 과정을 한번 더 거치게 된다. 이 과정에서 범주화에서 사용될 자질이 추가되는 것이다.

3.2 실험결과 분석

웹 문서를 전처리하여 범주화 한 실험집단의 분류 정확률은 72%로 전처리 하지 않은 집단의 분류 정확률(62%)보다 높게 나타났다. 컨텐트 문서는 전처리를 하지 않는 문서 유형이므로 결국 인덱스 문서를 전처리한 것이 범주화 성능을 향상시킨 요소가 된 것이라고 할 수 있다(표 3 참조).

문서 유형별 전처리 현황		최종 분류정확률
컨텐트문서	인덱스문서	
전처리 無	전처리 有	72%
	전처리 無	62%

< 표 3 > 문서 유형별 전처리 현황 및 최종 분류 정확률

인덱스 문서의 전처리는 인덱스 문서가 자동분류에 충분한 정보를 가지고 있지 못하다는 가정 하에 인덱스 문서가 내포하고 있는 정보, 즉 최종적으로 전달하고자 하는 컨텐트 문서로 인덱스 문서의 내용을 보강하는 방식이었는데, 이러한 시도가 분류 정확률을 개선시킨 것으로 나타났다.

또한 문서 유형간의 범주화 성능을 비교해보면 컨텐트 문서가 가장 범주화 성능이 좋은 문서 유형으로 판명되었다(표 4 참조). 컨텐트 문서의 분류 정확률은 전처리 하지 않은 인덱스 문서의 분류 정확률의 2배에 이르는 큰 차이를 보였다. 이 결과는 인덱스 문서가 적합한 클러스터링 형성을 저해하는 요소라고 제시한 컨텐트/인덱스 문서 클러스터링 실험 결과를 뒷받침하는 것이라 할 수 있다.

문서유형	분류 정확률
컨텐트 문서	81%
전처리한 인덱스 문서	63%
전처리 하지 않은 인덱스 문서	43%

< 표 4 > 문서 유형별 분류 정확률

확장한 인덱스 문서와 확장하지 않은 인덱스 문서에 할당된 범주를 좀 더 구체적으로 분석한 결과 몇 가지 특성이 나타났다. 인덱스 문서를 확장했을 때와 확장하지 않았을 때 모두 동일한 범주가 할당되었다면 해당 범주가 적합할 확률은 84%로 분류 정확률 측면으로 보자면 상당 수준에 이른다고 할 수 있다. 확장하지 않았을 때 적합한 범주를 할당받은 인덱스 문서는 확장한 경우에도 적합한 범주를 유지하는 확률이 높다고 할 수 있다. 이는 문서의 확장에 따른 내용변질이 크게 일어나고 있지 않다는 것이다. 또한 서로 할당된 범주가 다른 경우에는 확장한 문서에 할당된 범주가 적합할 확률(48%)이 확장하지 않은 문서에 할당된 범주가 적합할 확률(12%)의 4배에 달한다. 이 결과는 원래의 인덱스 문서가 적합하지 못한 범주를 할당받았을 경우, 인덱스 문서를 확장하면 원래 할당받았던 부적합한 범주를 적합한 범주로 변경시켜 줄 수 있는 효과가 있다는 것을 말해준다.

4. 결론

다양한 웹 문서들은 동일한 환경에서 범주화

하는 것 보다 문서의 특성을 고려하여 최적화 된 환경에서 각각 분리하여 범주화를 하는 것이 효과적이라는 가설을 실험하여 본 결과, 문서 유형별로 처리를 다르게 하여 범주화 한 경우가 더 효과적인 것으로 검증되었다. 컨텐트와 인덱스로 구성된 실험집단의 경우, 컨텐트와 인덱스에 각각 다른 방식을 적용하여 전처리를 한 범주화 결과가 전처리 하지 않은 방식보다 분류 정확률이 10% 향상되었다.

결과적으로, 확장한 인덱스 문서가 확장하지 않은 문서보다 범주화에 적합하였다고 할 수 있다. 그러나 인덱스 문서 범주화의 전체 정확률이 컨텐트에 비하여 낮으므로 이를 컨텐트 수준으로 향상시키는 것이 웹 문서의 범주화 성능을 개선하는데 있어 주요 관건이 될 것이다. 인덱스 문서 전처리 과정을 좀더 정련하여 처리 과정을 세분하는 것도 하나의 개선 방안이 될 수 있다.

또한 원래 인덱스 문서에 할당된 범주와 확장한 인덱스 문서에 할당된 범주가 다른 경우, 확장한 인덱스 문서에 할당된 범주가 맞을 확률이 원래의 인덱스 문서보다는 4배 가까이 향상되었지만 전반적으로 높은 편은 아니었으므로 역시 후속 연구가 필요한 부분이라 할 수 있다. 이 부분이 문서확장을 통한 인덱스 문서 범주화 성능 향상에 직접적인 영향을 미칠 수 있기 때문이다. 원 문서와 달리 범주가 할당되는 인덱스 문서만을 따로 처리하는 등 여러 다양한 기법을 시도해 보는 것이 바람직할 것이다.

본 연구에서 제안한 웹 문서 범주화 모형은 ①범주화 이전 단계에서 문서의 유형을 판단하고 그에 따라 전처리 하는 과정과 ②문서 범주화 과정의 2단계로 분리되어 있으므로, 웹 문서의 유형이 변화하거나 추가되더라도 유연하게 적용할 수 있다. 즉, 새로운 웹 문서 유형이 나타나는 경우 범주화 모형을 크게 수정하지 않고서도 해당 문서 유형에 대한 새로운 전처리 과정을 구축하여 쉽게 반영시킬 수 있는 가능성이 있는 것이다. 예를 들면 본 실험에서 제외시켰던 외부 링크 모음의 경우 핵심 구성요소인 링크의 URL을 범주화 자질로 선정하는 전처리 과정을 개발하여 본 모형에 쉽게 추가할 수 있다. 이는 범주화 대상인 웹 문서의 유형이 지속적으로 변화하고 발전하는 환경을 고려한 것이다.

참고문헌

조광제. 1997. **계층적 분류 구조상에서의 역카테고리 빈도를 이용한 문서의 자동 분류.** 석사학위논문, 동국대학교 대학원, 컴퓨터공학과.

Lewis, David D. 1992. *Representation and Learning in Information Retrieval.* Ph.D. diss., University of Massachusetts.

Parolli, Peter, James Pitkow and Ramana Rao. 1996. "Silk From a Sow's Ear: Extracting Usable Structure from the Web", *SIGCHI '96 Electronic Proceedings.*