웹 문서의 토픽 선정 방법에 관한 연구

공현장° 황명권 김판구
조선대학교 컴퓨터공학과
{kisofire°, mghwang, pkkim}@chosun.ac.kr

Study on the Topic Selection of Web Documents

Hyunjang Kong° Myunggwon Hwang Pankoo Kim
Dept. of Computer Engineering, Chosun University

요 약

웹 문서의 수가 기하급수적으로 늘어나는 현 시점에서 문서의 효율적인 관리을 위한 문서 클러스터링 방법은 현재 가장 요구되는 기술이다. 지금까지 문서 클러스터링의 방법 연구에서는 TF-Idf 측정값을 이용한 문서분류, Title 기반의 문서분류등과 같은 다양한 시도가 있었다. 이러한 문서 클러스터링 방법에서는 문서의 내용에 치중하거나 문서 분류를 위한 정확한 기준이 없어, 효율적인 문서의 클러스터링과 검색을 지원하지 못하였다. 그리하여, 본 연구에서는 새롭게 토픽 선정 알고리즘을 제안하고, 토픽 선정 알고리즘에 의해 결정된 토픽에 기반하여 문서 검색을 수행함으로써, 문서검색의 성능을 높일 수 있었다.

1. 서 론

오늘날 방대한 양의 웹 문서들이 존재하고 있다. 그리고 이러한 문서들에 대한 검색은 현재 검색 엔진에서는 문서의 제목이나 문서내의 키워드와 같은 몇가지 한정된 방법들을 사용하여 이루어지고 있다. 그렇지만 이러한 문서 검색의 전형적인 특징은 많은 결과값을 제공하는 반면, 검색 정확도가 상당히 낮다는 것이다. 이에 본 연구에서는 이러한 문서 검색의 효율성을 높이기 위하여 새로운 형태의 문서 검색의 방법론을 제안한다. 이러한 문서 검색의 성능 향상을 위해서 본 연구에서는 문서의 토픽을 효율적으로 선정하기 위한 방법론을 제안하며, 이렇게 효율적으로 선정된 토픽에 기반하여 문서검색을 수행함으로써, 좀더 효율적인 문서 검색을 기대할 수 있다. 본 연구에서는 문서내 토픽의 효율적 선정을 위하여 워드넷내 단어들사이의 계층 구조와 TF 측정 방법을 병행하여 토픽 선정 알고리즘을 구성하였다. 본 알고리즘을 사용함으로써 기계가 스스로 문서의 내용을 파악하고 문서내에서 가장 중요한 단어를 선정하여 문서의 토픽으로 선정하게 된다.

본 논문의 2장에서는 본 연구의 관련 연구를 소개하고, 3장에서는 토픽 선정을 위한 제안 알고리즘을 설명한다. 그리고 4장에서는 토픽 선정에 관한 실험 및 평가의 내용을 기술하고, 끝으로 결론 및 향후 연구 방향을 5장에 제시한다.

2. 관련연구

본 연구에서 토픽 결정을 위한 관련연구는 워드넷과 TF 측정 방법이다. 워드넷은 가장 대표적인 온톨로지 형태의 지식베이스이며, 이를 이용하여 오늘날 많은 연구들이 진행되고 있다.[5] 본 연구에서도 워드넷내의 단어들 사이의 연관성과 단어들 사이의 상하위 관계를 적용하여 문서의 토픽을 결정 하고자 하며, 또한 TF 측정값을 함께 사용하여 보다 효율적인 토픽 선정을 꾀한다.

2.1 WordNet

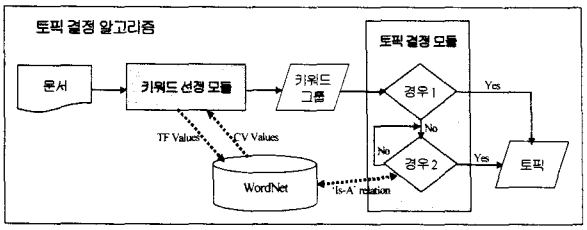
워드넷은 현재까지 가장 널리 사용되는 범용의 대형 온톨로지로써, 그 내용은 6개의 데이터베이스 테이블들로 구성되어져 실세계에 존재하는 어휘에 대해서 체계적으로 정의하고 있다. 워드넷은 크게 4개의 카테고리(명사, 형용사, 부사, 동사)로 분류되고, 그 안에는 다시 45개의 소카테고리로 분류되어져 있다. 그리고 워드넷내의 모든 개념들은 특정의 심볼들을 사용하여 각 개념들사이의 관계를 표현하고 있으며, 본 연구에서는 이러한 관계들 중에서 특히 개념들사이의 상하위 관계를 이용한다.

2.2 TF 측정 방법

대부분의 검색 엔진은 문서의 제목뿐만 아니라 문서 전체 내용을 인덱싱하는 방법을 사용한다. 인덱스의 방법을 사용할때는 관련있는 문서들을 링크로 연결한다. 이때, 하나 이상의 문서가 서로 연관이 있을 경우, 어느 문서가 사용자 질의에 더 관련있는 문서인지 구별할 필요가 있다. 이러한 순위를 결정하는데 가장 많이 사용되는 것이 TF-IDF 측정 알고리즘이다. TF는 한 단어가 한 문서내에 발생하는 횟수를 나타내고, DF는 한 단어가 N개의 문서 집합내에서 몇 개 문서에서 나타나는 횟수이다. 특정 단어가 한 문서내에 많이 나타난다면, 그 문서는 해당 단어에 대해서 중요한 문서라고 판단될 수 있지만, 여러개의 문서에서 고루 분포하여 나타난다면, 그 단어의 중요도는 떨어진다. 그리하여 문서에서 단어의 우선순위는 해당 문서에서 단어의 출현 빈도수와 역문헌 빈도의 곱으로 나타낸다.

3. 토픽 선정 알고리즘

본 연구에서 제안하는 토픽 선정 알고리즘은 두 과정을 수행하여 토픽을 결정한다. 첫 번째 과정은 문서내에서 주요 키워드들을 선정하는 과정이며, 두 번째는 선정된 키워드들에 기반하여 워드넷의 상하위 계층 구조를 이용하여 최종적으로 문서의 토픽을 결정하는 과정이다. 이러한 두 과정을 포함하고 있는 전체 토픽 선정 알고리즘의 수행 절차를 살펴보면 [그림1]과 같다.



[그림 1] 토픽 선정 알고리즘

[그림 1]에서와 같이, 본 알고리즘은 내부적으로 두 가지의 중요한 모듈을 포함하고 있으며, 이는 각각 키워드 선정 모듈과 토픽 결정 모듈이다. 또한 각각의 단계에서 워드넷에 접근하여 워드넷의 계층 구조 및 단어들 사이의 관계값을 적용하고 있다.

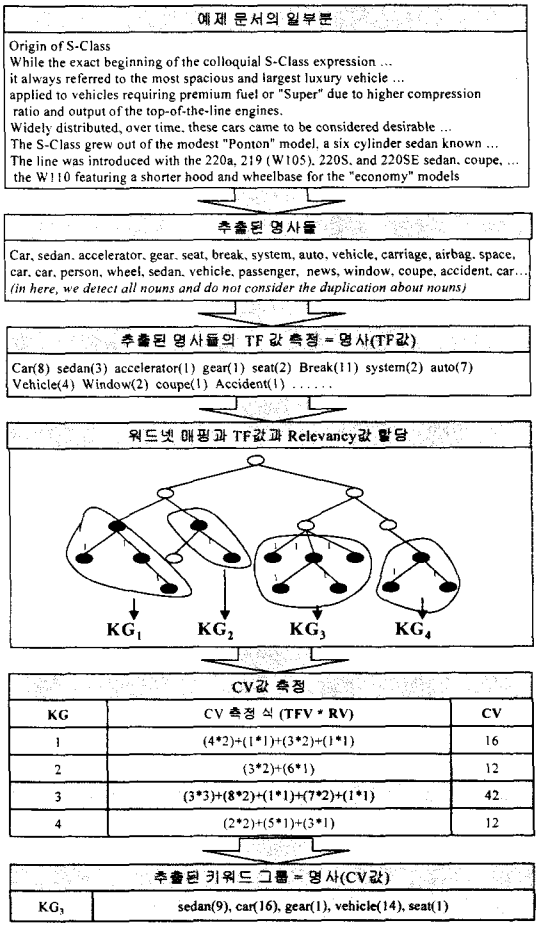
3.1. 키워드 선정 과정

먼저, 키워드 선정은 최종적으로 결정되는 토픽에 크게 영향을 미치는 중요한 모듈이다. 다시 말해서, 첫 번째 과정인 키워드 선정에서 선정된 키워드의 정확성 여부에 따라 토픽 결정의 정확도가 결정되어진다. 본 연구에서 제안하는 키워드 선정 방법은 다음과 같다. 기존의 대부분의 키워드 선정 방법에서 주로 단어의 출현 빈도수를 고려한 TF(Term Frequency)를 측정하여 키워드를 선정하여 왔다. 그렇지만 본 연구에서는 좀더 정확한 키워드의 선정을 위하여 TF값 측정과 더불어 단어들 사이의 관계성을 함께 고려하여 키워드 선정을 수행한다. 먼저 TF 값은 문서내의 단어들을 추출한 후, 단어의 출현 빈도를 계산하여 구할수 있으며, 이렇게 추출된 단어들은 워드넷내의 단어들과의 매칭을 통하여 관련된 단어들끼리 그룹화되고, 그룹화된 단어들 사이의 관계값을 고려하여 최종적으로 문서의 키워드 그룹이 선정된다. 키워드 선정 과정을 절차적으로 살펴보면 아래와 같다.

1. 문서내 단어 추출
2. 추출된 단어들의 TF 값 계산
3. 추출된 단어들을 워드넷내의 용어들에 맵핑
4. 맵핑된 단어들의 그룹화(워드넷내의 단어들사이의 관계 고려)
5. 그룹화된 단어들 사이의 관계값 측정
6. 최종적으로 키워드 선정을 위해 수식(1)을 사용하여 키워드 그룹을 선정

$$CV_i = \sum Ns_i(TFV_i * RV_i)$$
$$KWs = KG_i(Highest(CV_i)) \quad (1)$$

아래의 [그림 2]는 키워드 선정 모듈을 통하여 실제 문서내의 키워드 그룹을 선정하는 과정을 설명하고 있다.



[그림 2] 키워드 선정 모듈

[그림 2]에서 먼저, 웹 문서에서 명사들을 추출하고, 추출된 명사들의 TF 값을 계산한다. 그런다음, 명사들은 워드넷내의 단어들과 매칭되며, 매칭된 단어들은 워드넷내에서 관계성을 고려하여 관계가 있는 단어들은 그룹화된다. 이때 단어들 사이에 관계가 있을 때는 단어 사이에 관계값 1을 할당한다. 끝으로, 수식(1)을 적용하여 최종적으로 토픽 결정 모듈에서 사용할 키워드 그룹을 선정한다.

3.2. 토픽 결정 과정

3.1의 키워드 선정 과정을 거쳐 선정된 키워드들을 사용하여 토픽 결정의 과정에서는 최종적으로 문서의 토픽

을 결정하게 된다. 특히 토픽 결정 모듈에서는 두 가지의 경우를 고려하여 토픽을 결정하게 되는데, 첫 번째는 Specific 토픽의 결정을 위한 경우와 두 번째는 General 토픽 결정의 두 경우이다.

1. Specific 토픽 결정의 경우
2. General 토픽 결정의 경우

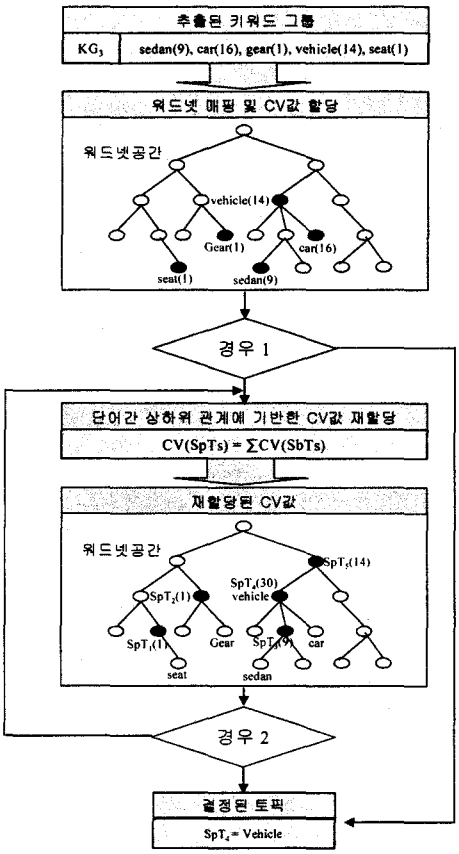
Specific 토픽의 결정의 경우는 키워드 선정 모듈에서 선정된 키워드들과 각각의 키워드들이 가지고 있는 CV 값을 고려하여 토픽을 결정한다. 이 경우에서는 키워드 그룹내의 단어들 중 CV 값을 고려하여, 다른 키워드들보다 상당히 큰 CV 값을 가지는 단어를 그 문서의 토픽으로 결정한다. Specific 토픽 결정을 위하여 다음의 수식(2)를 사용한다.

$$IF(CV(KW_i) > \sum CV(KW_i)^c$$
$$TOPIC = KW_i \qquad (2)$$

General 토픽 결정의 경우는 Specific 토픽 결정의 경우를 통하여 토픽 결정이 실패하였을 경우, 토픽의 결정을 워드넷내 단어들 사이의 상하위 관계를 고려하여 좀 더 General 한 토픽을 결정하는 방법이다. 이러한 과정에서 결정된 토픽은 문서에 대한 정확한 토픽은 아니지만, 대부분 비슷하거나 관련있는 토픽들이 결정된다. General 토픽 결정 과정에 대하여 자세하게 설명하면 다음과 같다. 우선, 키워드 그룹내 단어들을 워드넷 단어들과 매칭을 시킨후, 각 단어들이 가지고 있는 CV 값을 함께 포함할 수 있도록 워드넷의 내용을 새롭게 셋팅한다. 이때, 키워드 그룹에 존재하지 않으면서, 워드넷에 존재하는 단어들은 CV 값을 임의로 1로 셋팅을 한다. 그런 다음, 최하위 단계에서부터 순차적으로 단어들 사이의 상하위 관계를 조사한다. 그리고 상위의 단어는 하위 단어의 CV 값을 상속받게 되며, 상속받은 단어들 중에서 수식(3)을 만족하는 단어를 찾는다. 워드넷 내에서 단어들 사이의 상하위 관계의 의미는 상위의 단어가 하위 단어보다 좀더 일반적인 의미의 단어이므로, 이를 본 모듈에서는 General 토픽 결정이라 한다.

$$CV(SpTs) = \sum CV(SbTs)$$
$$IF(CV(SpT_i) > \sum CV(SpT_i)^c)$$
$$TOPIC = SpT_i \qquad (3)$$

만약 한번의 CV 값의 상속 절차를 수행한 후, 이를 만족하는 단어가 존재하면, 그 단어는 문서의 토픽이 되며, 존재하지 않는다면, 상위의 단어로 이동하면서 반복적으로 CV 값을 상속하여 상속된 단어의 CV 값에 대하여 수식(3)을 만족하는 토픽을 찾는다. 아와 같은 두 가지의 경우를 체크하여 본 연구에서 제안하는 토픽 결정 모듈은 구성되어져 있으며, 단어들 사이의 관계를 고려함으로써, 단순 단어의 빈도수를 고려하였을때의 토픽 결정의 낮은 정확률을 최대한 보완하고자 하였다. [그림 3]은 [그림 2]에서 선정된 키워드 그룹을 이용하여 토픽이 결정되는 과정을 설명하고 있다.



[그림 3] 토픽 결정 모듈

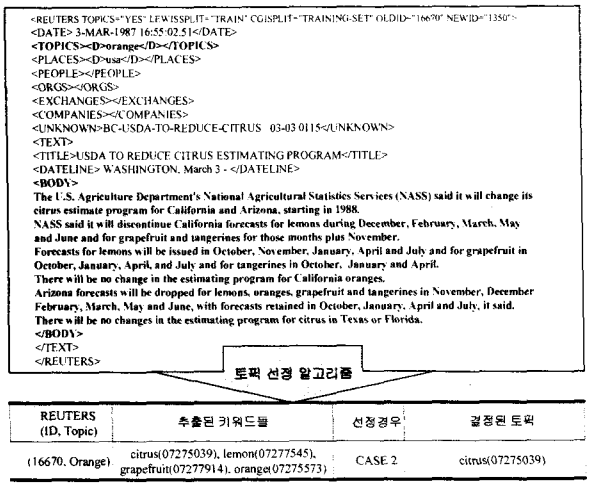
[그림 3]에서 먼저 키워드 그룹내의 단어들을 워드넷에 매칭시킨 후, 단어들에 대한 CV 값을 셋팅한다. 그런 다음, Specific 경우에 만족하는 토픽 결정을 시도한다. [그림 3]에서는 Specific 경우를 만족하지 않으므로, 토픽 결정이 실패하였고, 다시 General 토픽 결정 경우에 의해 단어들 사이의 상/하위 관계를 고려하여, 각 단어들의 CV 값을 재할당한 후, 수식(3)을 만족하는 토픽 결정을 시도하였으며, 한번의 CV 값 상속이 수행된 후, SpT_4가 최종적으로 문서의 토픽으로 결정되었다.

본 연구의 토픽 결정 알고리즘에서 문서의 토픽이 결정되면, 결정된 토픽에 기반하여 문서 검색을 수행하여, 문서 검색의 정확률을 높혀 줄 것으로 기대한다.

4. 실험 및 평가

앞에서 설명하였듯이, 본 알고리즘은 두 과정을 거쳐 토픽을 결정한다. 첫 번째 과정은 문서내에서 주요 키워드를 선정하는 과정이며, 두 번째는 선정된 키워드들에 기반하여 워드넷 단어들 사이의 상/하위 구조를 이용하여 최종적으로 토픽을 결정하는 과정이다. 본 알고리즘의 평가를 위해 다음의 실험을 수행하였다. 수행한 실험

의 내용은 [그림 4]에서 자세하게 설명하고 있다.



[그림 4] 실험 및 결과

[그림 4]에서 웹 문서는 REUTERS 데이터 중에서 orange 주제의 문서을 가지고 실험을 수행하였으며, REUTERS 문서는 그 내부적으로 토픽<TOPIC> 태그 형식으로 사람에 의해서 결정된 문서의 토픽을 표현하고 있다. 본 실험에서는 사람에 의해 결정된 토픽과 기계에 의해 결정된 문서의 토픽을 비교하는 실험을 수행하였다. 먼저, 실험 문서에 대한 키워드 선정 모듈을 통하여 키워드 그룹이 선정되며, 키워드 그룹내의 단어들은 각각 CV 값이 존재한다. 이렇게 선정된 키워드를 이용하여 토픽 결정 모듈에서 최종적으로 토픽이 결정되며, 본 실험에서는 General 토픽 결정의 경우에서 citrus가 문서의 토픽으로 결정되었다. 이는 본 연구이 토픽 결정 모듈을 통하여 정확하게 orange가 토픽으로 결정된 것은 아니지만, 의미적으로 좀더 General한 orange와 관련있는 citrus가 문서의 토픽으로 결정되었다. 실험에서도 알 수 있듯이, 문서에 대한 정확한 토픽이 선정되지는 않았지만, 토픽과 근접한 단어가 토픽으로 결정됨으로써, 의미적 토픽 결정이 이루어졌다고 할 수 있겠다. 이로써, 사람에 의해 수행되어져 왔던 막대한 데이터의 토픽 결정이 기계에 의해 이루어질 수 있을 것으로 기대된다.

5. 결론

결론적으로, 본 연구를 통하여 문서의 토픽 결정이 의미적으로 이루어짐을 확인할 수 있었으며, 이는 오늘날과 같은 막대한 양의 문서의 관리 및 검색의 분야에서 사람들에 의해 이루어지는 내용을 기계가 처리함으로써 효율적인 문서 관리와 검색을 기대할 수 있으며, 또한 문서의 처리 속도면에서도 사람에 의해서 처리되어질 수 없는 막대한 양에 대하여, 기계가 대신 처리하게 됨으로써 시간적 처리 비용이 절감될 것으로 기대된다. 그렇지만 본 연구에서의 실험 내용은 단순 문서에 대한 실험을 수행하여 그 신빙성이 조금 떨어진다. 향후, 다양한 문서에 대하여 실험을 수행하여 본 알고리즘의 타당성 입증을 위한 연구가 요구된다.

Acknowledgement

"본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음" (IITA-2006-C1090-0603-0040)

참고문헌

[1] Rauber, A. and Muller-Kogler, A. Integrating automatic genre analysis into digital libraries. In First ACM-IEEE Joint Conf on Digital Libraries. (2001).
[2] Sebastiani, F. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1-47. (2002).
[3] Y Yang, S. Slattery and R. Ghani. A study of approaches to hypertext categorization. Journal of Intelligent Information Systems, Volume 18, Number 2, March (2002).
[4] Kristina Toutanova, Francine Chen, Kris Popat & Thomas Hofmann. Text Classification in a Hierarchical Mixture Model for Small Training Sets, Xerox PARC and Brown University, Proceedings 10th International Conference on Information and Knowledge Management, (2001).
[5] http://wordnet.princeton.edu/
[6] Denoyer, L. Gallinari P. Bayesian Network Model for Semi-Structured Document Classification, in (Campos et al, (2004).
[7] Denoyer, L., Wisniewski, G., Gallinari, P. Document Structure Matching for Heterogeneous Corpora. In: SIGIR 2004, Workshop on Information Retrieval and XML. Sheffield, UK. (2004).