개념 기반 키워드 정보를 이용한 웹 문서의 자동 분류

박사준° 김기태
중앙대학교 컴퓨터공학과
{phdjoon°, ktkim}@ailab.cse.cau.ac.kr

Automatic Classification of Web Documents Using Concept-Based Keyword Information

Sajoon Park° Kitae Kim
Dept. of Computer Science and Engineering, Chungang-Ang University

요 약

본 연구에서는 웹 문서를 분류하기 위해서 분류하고자 하는 영역(category)에 대한 개념 지식을 이용한다. 먼저, 영역별 개념 지식을 기구축된 웹 문서의 집합으로부터 제목과 하이퍼링크에 기반한 앵커 텍스트를 이용하여 개념을 보유한 키워드를 추출한다. 추출된 키워드를 형태소 분석을 통해 색인어로 추출한다. 추출된 색인어에 대해 TFIDF를 확장한 영역 적용 색인 가중치 TFIDFc를 적용하여 영역별 개념 기반 색인어와 색인를 구축한다. 색인은 TFIDF를 영역별로 확장하여 구축한다. 구축된 영역별 개념 기반 색인을 이용하여 새로운 웹 문서에 대해서 어떤 영역에 해당하는 가를 결정하는 자동 분류 알고리즘을 수행한다. 자동 분류 알고리즘에 의해 수행된 문서는 영역별로 정리되며, 또한, 분류된 웹 문서의 색인어는 새로운 개념 기반 키워드로 추출되어 개념 기반 영역 지식을 구축한다.

1. 서론

정보 검색 방법에는 키워드 검색과 디렉토리 검색이 주로 이용된다. 디렉토리 검색을 사용하기 위해서는 먼저, 웹 문서를 디렉토리 별로 분류하여 보관할 필요가 있다. 이 때, 웹 문서를 디렉토리 별로 정확하게 분류, 보관 방법이 검색 결과의 정확도를 보장해 준다. 이제는 단순히, 검색 결과의 양보다는 정확성에 기준을 둔 질적인 면도 중요시 되고 있다. 디렉토리 별 분류 작업을 하기 위해서, 아직도 분류자가 수작업으로 웹 문서를 찾아서 분류 항목에 웹 문서를 삽입한다. 본 연구에서는 웹 문서를 분류하기 위해서 분류하고자 하는 영역(category)에 대한 개념 지식을 이용하고자 한다. 먼저, 영역별 개념 지식을 구축한다. 영역별 개념 지식을 구축하기 위해서는 영역을 대표할 만한 키워드를 이용하여 구축한다. 여기에서 사용되는 키워드는 제목과 하이퍼링크에 기반한 개념을 보유한 키워드를 사용한다. 전송 받은 웹 문서에서 제목, 하이퍼링크의 앵커 텍스트 단어를 추출하여 형태소 분석을 통한 색인어를 추출한다. 추출된 색인어를 이용 확장된 TFIDF를 적용 영역별 개념 기반 지식 DB를 구축한다. 구축된 영역별 개념 기반 색인을 이용하여 새로운 웹 문서에 대해서 어떤 영역에 해당하는 가를 결정하는 자동 분류 알고리즘을 수행한다. 자동 분류 알고리즘에 의해 수행된 문서는 영역별로 정리되며, 또한, 분류된 웹 문서의 색인어는 새로운 개념 기반 키워드로 추출되어 개념 기반 영역 지식을 구축한다.

본 논문에서는 2장에서 기반 연구를 살펴보고, 3장에서는 개념 기반 영역 문서 분류, 4장에서는 시스템 구성 및 실험을 수행한다. 마지막 5장에서는 결론 및 향후 과제를 제안한다.

이 논문은 2002학년도 중앙대학교 학술연구비 지원에 의한 것임

2. 기반 연구

2.1 문서 분류 방법

문서 분류란 정해진 분류 카테고리에 새로운 문서들을 가장 적합한 영역(category)에 배정하는 것이다. 일반적으로 문서내의 단어들을 이용하는 통계적인 방법이 사용되며, 많이 사용되는 통계적인 방법으로는 벡터 유사도를 이용하는 방법과 베이지안 확률을 이용하는 방법이 있다.

벡터 유사도(Vector Similarity)를 이용한 분류 방법은 분류하려는 문서와 각 분류 카테고리를 모두 단어들의 벡터 형태로 표현하고, 이들 두 벡터 사이의 각으로 유사도를 계산하여 가장 높은 유사도를 갖는 카테고리로 분류하는 방법이다[1].

베이지안 확률(Bayesian Probability)을 이용한 분류방법은 분류하려는 문서에 단어가 나타났을 때, 각 단어가 나타나는 사건이 독립적이라고 가정하면, 이 문서가 분류 카테고리에 분류될 확률은 문장에 속해 있는 용어들과 카테고리와의 결합 확률값에 대해 베이지안 규칙을 사용하는 방법이다[2].

2.2. 자동 색인(Automatic Indexing)

색인은 본문 중의 중요한 항목, 술어, 인명, 지명 등을 뽑아 한 곳에 모아, 이들의 본문에 위치한 페이지를 기재한 것으로, 인덱스 또는 찾아보기라고도 한다. 즉, 해당 문서를 표현하는 대표적인 단어들을 추출해 내어 기록하는 것이다.

과거에 색인 작업은 전문 사서에 의해 수행되었으나 현재에는 웹 문서에 대해 컴퓨터를 사용하여 해결하려는 자동 색인 기술이 연구되고 있다[3].

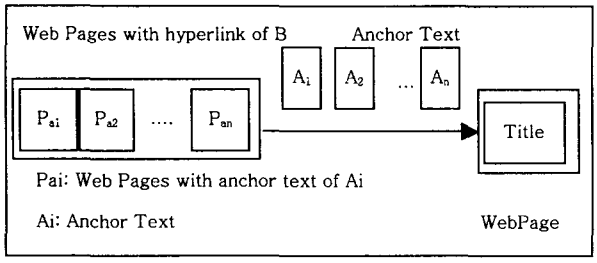
자동 색인은 문서에 포함되어 있는 색인어를 추출하는 과정으로 형태소 분석에 의하여 정보 자료를 분석하여 색인어 후보를 생성하는 과정과 생성된 후보 중에서 불용어 처리와 특수 색인어 추출 과정에 의하여 색인어를 선택하는 과정으로 이루어진다.

3. 개념 기반 영역 문서 분류

하이퍼링크는 현재 문서의 일부를 앵커 태그(<A>태그)를 이용해 하이퍼링크로 지정하고 이 링크가 현재 문서의 다른 부분 및 타문서를 가리키도록 한다. 이 때, 사용되는 단어나 구, 문장을 앵커 텍스트(anchor text.)라고 부른다.
하이퍼링크는 링크와 앵커 텍스트로 구성되며, 요약성, 연관성, 계층성, 보편성의 특성을 가진다.
웹 문서들 사이에 연결되어 있는 하이퍼링크의 연결 관계를 이용하면, 웹 문서들 사이의 상관 관계를 유추할 수 있다[4]. 이 때, 하이퍼링크들의 상관 관계를 특정 분야에 대한 웹 문서들로 한정을 하면, 하이퍼링크를 통하여 개념 기반 키워드를 추출할 수 있다.

3.1. 하이퍼링크를 이용한 키워드 추출

두 웹 문서가 하이퍼링크로 연결되어있다면 하이퍼링크는 두 웹 문서 사이의 관계를 앵커 텍스트를 통하여 나타낸다. 아래의 그림은 이 관계를 나타내 주고있다. 앵커 텍스트의 내용은 하이퍼링크의 특성에 따라 연결된 웹 문서의 키워드로 추출한다.

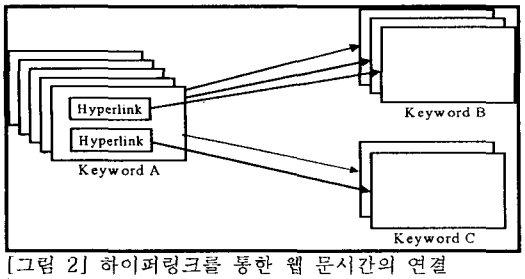


[그림 1] 핵심어 추출을 위한 구성 요소

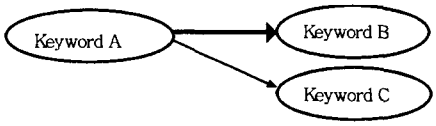
하이퍼링크를 이용한 개념 기반 키워드를 추출한 후 웹 문서는 하나 이상의 키워드를 가지게 된다. 각 문서의 키워드와 하이퍼링크인 링크를 이용하여 개념 관계를 생성한다. 링크는 계층적 구조 혹은 내용의 참조를 나타내는데, 이를 이용하여 웹 문서간의 계층적 구조와 참조가 키워드간의 계층적 구조와 참조로 추상화 되며, 이를 이용 개념 관계를 나타난다.

3.2 개념 기반 키워드 추출

특정 키워드를 가지는 웹 문서는 하나 이상이고, 웹 문서는 하이퍼링크 정보를 이용하여 서로 연결하고 있다. 특정 키워드를 가진 웹 문서가 연결하고 있는 웹 문서들의 키워드 리스트 중 동일한 키워드를 가지는 것끼리 분류를 한다 [그림2]. 그리고 키워드별로 분류된 문서들은 링크로 추상화하여 웹 문서간의 관계를 키워드간의 관계로 바꾸어 개념 관계를 생성한다[그림3].



[그림 2] 하이퍼링크를 통한 웹 문서간의 연결



[그림 3] 웹에서 추출된 키워드로 구성된 개념 관계

위의 과정을 거쳐 얻어진 키워드들간의 관계를 본 논문에서는 키워드간의 개념 관계라 한다.

3.3 영역 색인(category indexing)

개념 기반으로 추출된 키워드는 각 영역별로 색인화가 필요하다. 또한, 문서 분류시 측정 척도로 사용 될 가중치 평가 방법이 필요하다. 본 연구에서는 가장 보편적인 색인 가중치 방법 중 하나인 TFIDF 를 확장한다. TFIDF 는 하나의 문서 d 에서 단어 w 에 대한 가중치를 산출하는 방식이다. TFIDF 는 어떤 단어의 중요도는 그 단어가 문서에 나온 횟수(term frequency)에 비례하고, 그 단어가 있는 모든 문서의 총 수(document frequency)에 반비례한다는 의미이다[3].
본 연구에서는 한 문서가 아닌 한 영역별로 색인화가 수행되므로 문서에 대한 색인화를 영역으로 확장할 필요가 있다. 아래의 확장된 TFIDF를 이용하여 색인화를 수행한다.

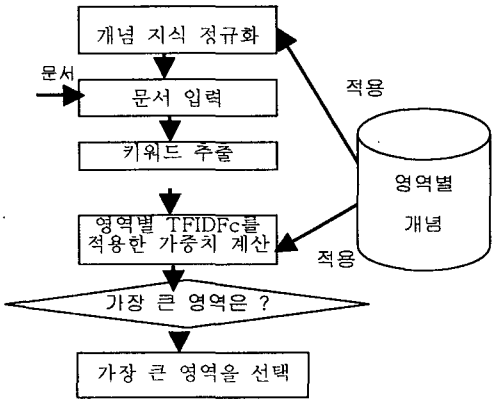
$$TFIDFc(k,C) = \sqrt{\mathrm{TF(k,C)} * \frac{N}{DF(k)}} \quad (1)$$

TFIDFc(k,C) : 영역 C에서 키워드 k에 대한 TFIDF
TF(k,C) : 영역 C에서 키워드 k가 출현한 횟수
DF(k): 영역 C에서 키워드 k가 출현한 문서의 총 수
N: 영역 C에 존재하는 문서의 총수

영역 C에서의 TFIDF의 가중치는 전체 영역에서 출현하는 키워드에 대해서 단어 출현 빈도와 문서 출현 빈도를 구하여 키워드의 색인화 작업을 수행한다. 영역에서 출현 빈도가 많은 단어가 영역에서 중요한 역할을 수행하고 출현 문서가 적을수록 키워드의 정보 가치가 높다고 볼 수 있다.

3.4 문서 분류

문서 분류는 웹 문서가 어떤 영역에 속하는 지를 결정해 주는 과정이다. 문서 분류 방법은 웹 문서 수집기와 인덱서를 통해서 구축된 개념 지식을 사용하여 입력 받은 문서를 영역별로 분류하는 역할을 수행한다.



[그림 4] 문서 분류 과정

우선, 문서를 분류 하기 전에 기구축된 영역으로부터 영역별 개념 지식 DB를 구축한다. 구축된 영역별 DB는 영역별 색인어의 크기에 차이가 있으므로 이를 정규화하는 과정을 수행한다. 정규화 과정 완료 후에 문서를 입력하여 영역별로 문서를 분류한다. 문서를 분류하는 방법은 입력 문서를 색인어 추출기에 입력하여, 입력된 문서의 키워드를 추출한다. 그런 다음 추출된 색인어를 각 영역의 개념 지식에 적용하여 문서의 키워드의 가중치를 위에서 정의한 TFIDFc를 이용하여 키워드의 총합를 계산한다. 계산 결과 최고치를 얻은 영역으로 문서의 영역을 결정한다.

4. 시스템 구성 및 실험

4.1 시스템 구성

시스템은 웹 로봇(web robot), 인덱서(indexer), 개념 지식을 보유한 영역별 DB, 문서 분류기로 구성된다.
웹 로봇은 다수의 클라이언트가 하나의 서버에 연결되어 서버의 지시에 따라 웹 문서를 수집한다. 인덱서는 수집한 웹 문서에서 색인어를 추출하여 인덱싱 작업을 하여 영역별 DB에 개념 지식을 구축한다. 영역별 DB는 수집된 문서로부터 추출된 개념 지식을 데이터베이스화 한다. 문서 분류기는 웹 문서 수집기와 인덱서를 통해서 구축된 개념 지식을 사용하여 입력되는 문서를 각 영역별로 분류한다.

4.2 실험 및 결과

4.2.1. 실험 환경

본 논문에서는 영역별 DB를 구축하기 위해서 대표적인 인터넷 포탈 사이트인 야후 코리아(Yahoo Korea)의 웹 디렉토리에서 각 영역의 웹 문서를 수집하였다. 수집한 영역은 야후 코리아의 비즈니스와 경제(http://kr.dir.yahoo.com/ business_and_economy/), 정부 (http://kr.dir.yahoo.com/government/), 엔터테인먼트 (http://kr.dir.yahoo.com/entertainment/) 등 3개의 웹 디렉토리이다. 시작 웹 문서에서부터 하이퍼링크가 5번 연결된 웹 문서까지 수집하였다. 수집한 웹 문서의 개수는 비즈니스와 경제 3000개, 정부 900개, 엔터테인먼트 2000개이다. 각 카테고리별로 크기의 차이가 생김을 알 수 있고 이를 위해 정규화 과정이 필요하다.

4.2.2 결과 및 평가

문서 분류를 실험하기 위하여 일간지에서 정치, 경제, 연예 섹션에서 신문 기사를 무작위로 추출하였다. 실험에 사용된 신문 기사는 경제, 정부, 연예 세 분야이다. 신문 기사의 개수는 총 500개이며, 경제 250개, 정부 120개, 연예 130개로 구성되어 있다.

표1. 실험에 사용된 신문 기사 영역별 분포

영역(Category)	실험에 사용되는 신문 기사의 개수
경제 (비즈니스와 경제)	250
정부	120
연예 (엔터테인먼트)	130
총 합	500

표2. 정확도 비교

분류 방법	잘못 분류한 문서 개수	정확도
단순 단어 빈도	120 개	76 %
단순 TFIDF	80 개	84 %
본 논문	60 개	88 %

실험 결과를 보면, 단순 단어 빈도를 이용한 문서 분류 방법이 76 %의 정확도로 가장 나쁜 성능을 보였으며, 단순 TFIDF를 이용한 문서 분류 방법이 84%의 정확도로 그 다음으로 좋은 성능을 보였다. 그에 반하여, 본 논문에서 제안한 개념 기반 문서 분류의 경우는 88%의 정확도를 보였다.
실험에 분류를 위해 사용한 영역은 경제, 정부, 연예이다. 경제와 정부는 서로 매우 밀접한 관계를 가지고 있다. 그래서 실험 결과를 보면 경제를 정부로 분류한 예가 많았다. 반면에 연예의 경우 경제나 정부와는 상대적으로 밀접하지 않기 때문에 신문기사를 정확히 분류했다.

5. 결론 및 향후 과제

기존의 연구에서는 개념 지식을 구축할 때 사람의 개입이 필요한 수작업이 많았다. 그래서 시스템 구축 과정에서 병목 구간으로 작용하였다. 그러나 본 논문에서는 사람의 개입 없이 개념 지식을 자동으로 구축하여 시스템 구축 과정의 병목 구간을 제거할 수 있는 방안의 개발을 시도하였다.
문서를 분류하기 위해서는 문서가 전달하고자 하는 주제의 의미(semantic)를 이해할 수 있어야 한다. 그래서 본 논문에서는 통계적 기법 뿐만 아니라 하이퍼링크 정보를 사용하여 연결 관계의 의미를 이용하여 문서를 영역별로 분류하는 시도를 하였고 필요한 알고리즘을 제안하였다.
문서를 분류하기 위하여 인터넷의 웹 디렉토리와 형태소 분석기를 이용한 색인어 추출기를 사용하여 색인어를 추출하였다. 그리고 추출된 색인어를 영역별 개념 지식으로 구축했다. 문서 분류기에서는 문서를 분류하기 위하여 앞에서 구축한 개념 지식과, TFIDF를 확장한 영역별 TFIDF를 이용 문서를 분류하였다.
본 논문에서 제안한 방법과 알고리즘을 이용 문서 분류시 88%의 정확도를 보여 문서 분류의 자동화에 대한 가능성을 증명하였다.
본 논문에서는 웹 문서의 제목, 하이퍼링크의 앵커 텍스트만을 이용 영역별 개념 지식을 구축하였으나 본문에 있는 키워드도 함께 사용하는 방법을 계속 연구하고자 한다. 또한, 추출된 개념 관계를 이용하여 키워드간 관계의 특징을 추출하는 연구가 필요하다.
과거부터 문서를 영역별로 분류하기 위한 많은 노력과 연구가 있었다. 또, 문서 안에서 문서가 가지고 있는 정보를 추출하여 요약, 색인화 하려는 많은 노력들이 있었다. 문서 분류의 자동화와 정확도 향상이 더 필요하다고 본다.

참고문헌

[1] W.Frakes and R. Baeza-Yates, Information Retrieval, Prentice Hall, 1992
[2] Lewis, David D. (1998). Naïve (Bayes) at forty: The indepedence assumption in information retrieval. Proceedings of ECML-98, 10th European Conference on Machine Learning, 1998.
[3] Salton, G., and Buckley, C., "Term weighting approaches in automatic text retrieval", Tech Report 87-881 Dept. of Computer Science, Cornell University, 1987
[4] 박사준, 김상경, 황수철, 김기태, "전문가 검색 엔진에서 개념 그래프를 이용한 web 정보 획득," 2000년 봄 학술발표논문집(B) 제 27권 1호 페이지295-297, 한국정보과학회, 2000.
[5]Sa-Joon Park, Jae-Ho Kim, Ki-Tae Kim, "WebExpert, Expert Search Engine On A Specialized Field Using The Conceptual Relationship," Proceeding of the Sixth IASTED International Conference INTERNET AND MULTIMEDIA SYSTEMS AND APPLICATIONS, pp.86-91, 2002