일관성 있는 문서분류 및 키워드 추출을 위한 말뭉치 구축도구

정재철 ·박소영 ·장준호 ·길태숙

상명대학교

A Corpus Construction System of Consistent Document Categorization and Keyword Extraction

Jae-cheol Jeong ·So-Young Park ·Juno Chang ·Tae-Suk Kihl

Sangmyung University

E-mail : ssoya@smu.ac.kr

요 약

최근에 웹 문서의 양이 빠르게 증가함에 따라 사용자가 원하는 정보를 검색하기 위한 효율적인 문서분류방법에 대한 연구가 요구되고 있다. 본 논문에서는 효율적인 문서분류 시스템 개발을 위한 자료수집 단계에서, 제시되는 각 문서에 대해 일관성 있는 문서범주 및 사용용도, 키워드 정보를 부착하기 위한 말뭉치구축 도구를 제안한다. 이 때 다른 사용자가 입력한 정보를 제시함으로써 자신의 것과 비교 및 수정할 수 있는 검증단계를 거쳐 일관성을 높인다. 또한 웹 환경에서 실행하여 말뭉치 구축자가 언제 어디서든지 편하게 말뭉치를 구축할 수 있다.

ABSTRACT

As the number of documents rapidly increases in the web environment, the efficient document classification approaches have been required to retrieve the desired information from too many documents. In this paper, we propose a corpus construction tool to annotate document classification information such as category, keywords, and usage to each product description document. The proposed tool can help a human annotator to correctly identify this information by providing the verification step to check the input results of other human annotators. Also, the human annotator can construct the corpus anytime anywhere by using the web-based proposed system.

키워드

말뭉치, 자동문서분류

I. 서 론

인터넷이 발달함에 따라 각종 게시판, 인터넷 신문, 블로그 등에서 수많은 문서들이 생성되고 있다. 또한 사용자는 다양한 문서들 중 자신이 관심 있는 문서를 찾기란 쉬운 일이 아니다.

현재 국내외에서 자동 문서분류 방법에 대한 연구가 활발히 이뤄지고 있고, 이들 대부분이 학습이나 통계 혹은 확률 자료에 의한 것이 주류를 이루고 있다.[1] 또한 기계에 의한 문서분류를 위해서는 범주 집합이 미리 주어져야하기 때문에[2] 기계학습을 위한 말뭉치 구축은 자동 문서분류 방법에 시작이라고 할 수 있고, 시스템의 성능을 좌우한다고 할 수 있을 정도로 말뭉치의 신뢰도는 중요하다. 말뭉치 구축시, 구축된 말뭉치에 필요한 정보가 포함되어 있는지 고려되어야하고, 그 정보는 일관성을 가져 신뢰성이 있어야 한다. 또한 말뭉치 구축자에게 편안한 환경을 제공해야하는지도 고려되어야 한다.[3]

따라서 본 논문에서는 다수의 말뭉치 구축자에게 웹 환경에서 언제 어디서나 편하게 문서분류 정보를 부착하도록 하고 말뭉치를 추출한 후 다른 구축자의 정보 부착 결과와 비교 및 수정하여 일관성 있고 신뢰도 높은 문서분류 말뭉치를 구축하는 시스템을 제안하고자 한다.

II. 제안하는 시스템

제안하는 시스템은 문서분류 정보와 키워드 정보를 포함하고 있는 말뭉치를 얻기 위한 사전 조사 시스템으로 말뭉치 구축자가 언제, 어디에서든지 웹 페이지에 접속하여 조사에 참여할 수 있는 특징이 있다. 또한 시스템은 각 말뭉치 구축자에게 자신이 입력한 정보뿐만 아니라 같은 문서에 대한 다른 구축자의 입력 정보를 제시하기 때문에 자신의 생각과 비교하여 재고할 수 있는 기회를 부여한다. 또한 말뭉치를 구축하는데 있어서 시험 문서의 수가 충분히 커야 정확한 정보를 얻어낼 수 있기 때문에 대량의 시험 문서가 필요하고[4], 이를 고려하여 현재 시험 중인 문서의 로그를 DB에 남겨두어 시스템에 다시 접속하더라도 진행 중인 문서부터 시작할 수 있게 하였다.

그림 1. 키워드, 사용용도, 범주 입력화면

그림 2. 입력정보 검증화면

시스템에 접속하여 로그인 하면 [그림 1]의 왼쪽과 같이 말뭉치 구축자에게 [그림 1]의 오른쪽과 같이 임의의 상품에 관한 설명 문서를 제시하고 해당 문서와 관련된 키워드를 자유롭게 입력받는다. 그리고 해당 문서의 사용용도에 대한 정보를 입력받은 후 사전에 준비한 문서범주를 제시하여 선택하게 한다. 말뭉치 구축자가 입력한 정보는 데이터베이스에 저장되고 모든 문서를 시험할 때 까지 같은 작업이 반복된다. 이전 단계 작업이 완료되면, [그림 2]와 같이 입력된 문서범주에 관한 검증작업 페이지로 이동한다. 페이지는 말뭉치 구축자에게 자신이 입력한 키워드, 사용용도, 문서범주 뿐만 아니라 다른 말뭉치 구축자가 입력한 정보까지 제시하고 이를 참고하여 말뭉치 구축자에게 재검토 및 수정의 기회를 제공한다. 또한 문서분류에 관련해 토론의 여지가 있을 경우 보류할 수 있는 기능을 제공해 차후에 판단하도록 한다.

모든 문서를 검토하면 키워드 추출을 위한 페이지가 제시된다. 시스템은 말뭉치 구축자에게 임의의 문서를 제시하고 말뭉치 구축자가 입력한 사용용도, 문서범주 정보와 시스템이 선별한 30개의 키워드를 제시받는다. 말뭉치 구축자는 제시받은 키워드 중 5개를 선택하게 되는데 키워드는 자신이 선택한 범주를 참조하여 이전 단계에서 해당 범주로 저장된 키워드를 탐색하여 선별된다. 모든 문서를 시험할 때 까지 같은 작업이 반복된다. 키워드 추출 단계도 이전 단계와 같은 검증단계가 진행된다. 말뭉치 구축자는 자신의 키워드 입력 정보뿐만 아니라 다른 사용자의 키워드 입력 정보를 함께 제시받고 자신의 의견을 재검토 및 수정할 기회를 받는다.

제안하는 시스템은 JSP를 활용하여 구현되었으며, MySQL기반의 데이터베이스를 구축하였고, Apache Tomcat 서버를 이용하였다.

III. 결 론

본 논문에서는 일관성 있는 문서분류와 키워드 추출을 위해 다음과 같은 특징을 갖는 말뭉치도구 시스템을 제안하였다.

첫째, 제안하는 시스템은 문서분류를 위한 말뭉치 구축도구로 말뭉치 구축자가 각 문서에 대해 분류정보, 사용용도, 키워드 정보를 부착할 수 있도록 지원한다.

둘째, 제안하는 시스템은 검증 단계에서 자신의 입력정보는 물론 다른 구축자의 입력정보를 참조할 수 있게 제시하여, 해당 문서와 관련하여 일관성 있는 말뭉치를 구축하게 하였다.

셋째, 제안하는 시스템은 인터넷 웹 페이지 기반 환경에서 구현되었기 때문에 다수의 구축자가 시간과 장소의 구애를 받지 않게 하여 편안한 구축환경을 제공하였다.

참고문헌

[1] 최동시, 정경택, “카테고리와 키워드의 밀접성 정보에 의한 문서 자동 분류 시스템 설계 및 구현”, 한국정보과학회 학술발표논문집, pp. 639~642, 1995

[2] 나동열, 김유식, 신현주, 김태규, 강현규, 최호섭, 윤화묵, "한국어 문서분류 테스트컬렉션 개발", 한국콘텐츠학회 종합학술대회 논문집, 제5권 2호, pp435~439, 2007

[3] 임준호, 곽용재, 박소영, 임해창 "신경망을 이용한 반자동 구문분석 말뭉치 구축도구", 한국정보과학회 2003년도 봄 학술발표 논문집, pp. 483~485, 2003

[4] 송만석, 양단희, “한국어 기계학습과 말뭉치 구축”, 한국정보과학회 학술발표 논문집, 제25권 1호, pp.408~410, 1998