

연관규칙을 이용한 문헌정보학 전문용어 클러스터링 기법에 관한 연구

A Clustering Technique Using Association Rules for The Library and Information Science Terminology

승 현 우(Hyon-Woo Seung)

박 미 영(Mi-Young Park)

초 록

본 논문에서는 대량의 웹 문서로부터 연관된 지식정보를 검색하기 위한 전문 검색엔진을 개발하기 위하여 텍스트에서 추출된 전문 용어를 효율적으로 클러스터링하기 위한 방법을 제안하고자 한다. 즉, 일반적인 용어들간의 무의미한 연관 규칙이 양산되는 것을 방지하기 위하여 전문 용어로 구성된 지식베이스 테이블을 이용하여 의미 있는 용어들간의 연관 규칙을 생성한다. 연관 규칙은 하나의 논문에서 사용된 전문 용어들의 집합을 트랜잭션 단위로 구성하여 Apriori 알고리즘을 적용하여 생성된다. 하나의 용어로부터 생성된 연관 규칙 집합은 해당 전문 용어와 관련된 클러스터로 구성된다.

ABSTRACTS

In this paper, an effective method for clustering terminologies extracted from text is proposed, in order to develop a search engine to extract relevant information from large web documents. To prevent frequency of the meaningless association rules among general terminologies, only useful association rules among terminologies are produced using database tables which consist of domain-specific terminologies. Such association rules are produced by applying the Apriori algorithm after forming transaction units from groups of association rules in a document. A group of association rules produced from a terminology forms in a cluster.

키워드: 데이터마이닝, 연관규칙, 클러스터링, TF-IDF, Apriori 알고리즘

Data Mining, Association Rules, Clustering, TF-IDF, Apriori Algorithm

본 논문은 서울여자대학교 2003년 교내특별연구과제지원에 의해 수행된 것임.

서울여자대학교 정보통신공학부 컴퓨터공학전공 교수(hwseung@swu.ac.kr)

서울여자대학교 정보영상학부 문헌정보학전공 강사(ollive@swu.ac.kr)

논문접수일자 2003년 5월 13일

게재확정일자 2003년 6월 15일

참고문헌

강승식, HAM : 한국어 형태소 분석 라이브러리.

<<http://nlp.kookmin.ac.kr/HAM/Kor/download.html>>

김호성, 고희정. 1992. 용어 빈도수를 이용한 영문 문헌정보의 점진적인 개념적 집단화. 『한국정보과학회논문지』, 19(1): 12-23.

박건호. 1999. 『Apriori 알고리즘 연관규칙 마이닝기법을 이용한 정보검색』. 석사학위논문, 고려대학교 대학원, 컴퓨터학과.

박정오, 황도삼. 2000. 전문 용어 추출 시스템. 『한국정보과학회』, 27(1): 316-318.

박종수, 유원경, 홍기형. 1998. 연관 규칙 탐사와 그 응용. 『한국정보과학회 SIGDB 춘계특토리얼』.

서성보, 김선철, 이준욱, 류근호. 2000. 주요 항목 집합을 이용한 문서 클러스터링 및 연관 규칙 탐사 기법. 『한국정보과학회』, 27(1): 169-171.

승현우, 박미영, 조용한, 강미나. 2002. 상용 DataMining Tool 비교 분석. 『제3회 한국정보처리학회 정보기술워크숍』

승현우, 박미영, 황정민. 2002. 데이터마이닝기법과 연구동향분석. 『제4회 한국정보처리학회 정보기술워크숍』

승현우, 박미영, 황정민. 2002. 데이터마이닝기법을 이용한 문헌정보학 전문용어 클러스터링 데이터베이스 설계에 관한 연구. 『서울여자대학교 자연과학연구소논문집』

신진섭. 2000. 『웹 문서 분류를 위한 단어의 연관성 모델과 클러스터링 모델』. 박사학위논문, 건국대학교 대학원, 컴퓨터정보통신공학과.

이란주. 데이터마이닝에 관한 연구.

<<http://203.241.185.12/asd/main.cgi?board=Data&ryal=&view=2&back=&search=%C0%CC%B6%F5%C1%D6&where=1&how=1>>

이문기, 권오욱, 이종혁. 2000. 웹 디렉토리 서비스를 위한 문서 클러스터링. 『한국정보과학회』. 27(1): 351-353.

이정원 외 12인. 2000. 『데이터마이닝 알고리즘 분석』. 이화여자대학교 과학기술대학원 컴퓨터학과 EIST Research Report Series.

이정화. 2000. 『데이터마이닝 기법을 이용한 전문용어 클러스터링』. 석사학위논문, 창원대학교 대학원, 컴퓨터학과.

Eui-Hong Han, Vipin Kumar, George Karypis. 1997. 『Hypergraph Based Clustering In A HighDimensional Data Sets : A Summary of Results』. IEEE.

Han, Jiawei. 2002. Data Mining: Concepts and Techniques. Canada: Morgan Kaufmann publishers.

R. Agrawal, Chu Xu. 1994. 『Fast Algorithms for Mining Association Rules』, In Proc. of VLDB.

Thorsten Joachims. 1996. 『A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization』. CMU-CS-96-18.

William B. Frakes, Ricardo Baeza-Yates. 1992. 『Information Retrieval : Data Structures & Algorithms』. Prentice Hall.

