https://doi.org/10.5392/JKCA.2017.17.11.389

클라우드 환경에서 문서의 유형 분류를 위한 시맨틱 클러스터링 모델

Semantic Clustering Model for Analytical Classification of Documents in Cloud Environment

김영수, 이병엽
배재대학교 사이버보안학과

Young Soo Kim(experkim@gmail.com), Byoung Yup Lee(bylee@pcu.ac.kr)

요약

최근 시맨틱 웹 문서는 클라우드 기반으로 생성 및 유통되고 문서유형 분류에 따른 쉽고 신속한 정보 검색을 위해 지능형 시맨틱 에이전트를 요구하고 있다. 기존의 웹 문서의 검색은 키워드를 이용하여 해당하는 질의어가 포함된 문서 목록을 결과로 가져오며 사용자의 요구시에 내용을 제시하는 것이 일반적인 형태이다. 이는 웹 문서의 유사도와 시맨틱 관련성을 고려하지 않음으로써 사용자가 내용 검색과 분석에 많은 시간과 노력을 요구한다. 이의 해결을 위해서 빅 데이터 요소 기술인 하둡과 NoSQL을 활용하여 시맨틱 웹 문서에 포함된 키워드 빈도에 기반한 웹 문서의 유형 분류와 유사도를 제시하는 시맨틱 클러스터링 모델을 제안한다. 제안 모델은 실시간 데이터 처리가 요청되는 이종 모델을 가진 공공 데이터와 웹 데이터를 취합하여 일반 사용자가 쉽게 질의할 수 있는 대용량 지식 기반 시스템을 구축하는데 응용 모델로 활용될 수 있다.

■ 중심어 : | 클라우드 | 시맨틱 | 키워드 빈도수 | 컨셉 매칭 | 클러스터링 |

Abstract

Recently semantic web document is produced and added in repository in a cloud computing environment and requires an intelligent semantic agent for analytical classification of documents and information retrieval. The traditional methods of information retrieval uses keyword for query and delivers a document list returned by the search. Users carry a heavy workload for examination of contents because a former method of the information retrieval don't provide a lot of semantic similarity information. To solve these problems, we suggest a key word frequency and concept matching based semantic clustering model using hadoop and NoSQL to improve classification accuracy of the similarity. Implementation of our suggested technique in a cloud computing environment offers the ability to classify and discover similar document with improved accuracy of the classification. This suggested model is expected to be use in the semantic web retrieval system construction that can make it more flexible in retrieving proper document.

■ keyword : | Cloud | Semantic | Keyword Frequency | Concept Matching | Clustering |

* 이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2017R1A2B1003678). 이 논문은 2017학년도 배재대학교 교내학술연구비 지원에 의하여 수행된 것임.

접수일자 : 2017년 08월 04일
수정일자 : 2017년 09월 11일
심사완료일 : 2017년 09월 11일
교신저자 : 이병엽, e-mail : bylee@pcu.ac.kr

Ⅰ. 서 론

클라우드 컴퓨팅은 비즈니스의 변화에 기민하게 대응하고 전체적인 비용을 줄여주면서 효율성을 높이기 위해서 전 세계적으로 분산되어 있는 리소스의 확장성, 신뢰성, 가용성과 같은 특징을 갖는 기술이다. 클라우드 컴퓨팅 서비스는 인터넷과 웹 기반으로 제공되는 서비스로 SaaS는 주문형 소프트웨어 배포 모델이다. 디지털 콘텐츠의 폭발적 증가로 인해서 클라우드 컴퓨팅은 SaaS 구현 애플리케이션인 콘텐츠관리시스템의 기반 기술로 활용되고 있다. 질의어를 포함한 웹 문서를 신속하게 발견할 수 있도록 시맨틱 웹이 현재의 웹의 확장으로 제안되고 있다[1-6]. 웹 문서 클러스터링에 대한 이전 연구들은 비구조적인 시맨틱 정보들을 고려하지 않고 있다. 시맨틱 웹은 웹문서를 쉽고 빠르게 찾고 상호작용하고 협업할 수 있는 환경을 제공한다. 웹 문서에 구조적 정보로서 Table이나 DIV 태그와 같은 정보를 포함시켜서 데이터에 대한 디스플레이 구조를 정의하고 시맨틱 정보로서 article과 footer와 같은 시맨틱 태그를 포함시켜 문서의 의미적 구조를 정의함으로써 정보 검색과 지식 발견을 용이하게 한다. 본 논문에서는 문서의 구조적 정보와 시맨틱 정보를 둘 다 고려하는 유사도 척도 기반의 시맨틱 클러스터링 기법을 제안한다. 클러스터링 과정은 질의어와 웹문서의 유사도를 평가하여 웹 문서를 그룹화 하는 과정으로 검색 결과의 브라우징을 위해서 목록의 형태로 문서 그룹을 구조화한다. 기존의 문서 클러스터링 방법인 검색어의 차원 표현에 제한을 받는 문제점을 안고 있는 반면에 제안 모델에서 사용하는 클러스터링은 K-Means 클러스터링 알고리즘과 키워드의 출현 빈도수와 컨셉 매칭을 사용하고 웹문서를 마이닝 하고 관리하기 위하여 클라우드 기반의 NoSQL과 하둡을 사용하였다. [그림 1]과 같은 연구 모델을 사용하여 클라우드 환경에서 콘텐츠의 유형 분류를 위한 시맨틱 컴퓨팅 모델을 제안한다.

본 논문은 다음과 같이 구성된다. 2장에서는 클라우드 기반의 시맨틱 웹 모델을 분석하였고 3장에서는 하둡 기반 클러스터링 모델을 분석하였다. 4장에서는 키워드 빈도수와 컨셉 매칭 기반의 시맨틱 클러스터링 모델을 제안하였다. 5장에서는 결론과 시사점을 기술한다.

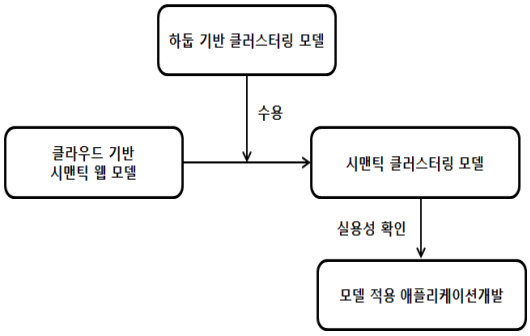

그림 1. 연구모델

Ⅱ. 클라우드 기반의 시맨틱 웹 모델

2.1 클라우드 컴퓨팅 모델

클라우드 컴퓨팅은 서버와 스토리지 그리고 애플리케이션 등과 같이 설정 또는 공유가 가능한 컴퓨팅 자원에 대하여 언제 어디서나 편리하게 접속이 가능한 기술이다. 클라우드는 추상화와 가상화를 통하여 사용자와 개발자가 시스템의 자세한 내용을 몰라도 이용거나 수정할 수 있고 공유되는 시스템 자원을 통해 다수의 사용자와 개발자들이 사용할 수 있는 기반 환경을 제공한다[7-10]. 사용자와 개발자가 이용 가능한 클라우드 컴퓨팅 서비스는 크게 3가지 형태로 분류한다. IaaS(Infrastructure as a Service)는 사용자가 가상 머신과 같은 하드웨어 자원을 사용할 수 있도록 서비스를 제공한다. IaaS는 기존 전통적인 시스템과는 다르게 하드웨어 사이에 가상화 계층이 존재한다. 가상화 계층은 하이퍼바이저를 주요 요소기술로 가지며 다수의 가상 머신이 구동될 수 있는 환경을 제공한다.

PaaS(Platform as a Service)는 사용자에게 운영체제와 프레임워크와 같은 플랫폼을 사용할 수 있도록 해주는 서비스로 이를 통해서 사용자는 애플리케이션을 개발하여 서비스를 제공할 수 있는 형태이다. SaaS(Software as a Service)는 애플리케이션 그리고 사용자 인터페이스를 포함하는 서비스로 사용자가 애

플리케이션을 개발해서 이용하는 형태가 아니라 클라우드 서비스 제공자가 애플리케이션을 개발하고 이를 사용할 수 있게 하는 형태의 서비스이다. 제안 모델에 적합한 클라우드 컴퓨팅 서비스는 SaaS로써 사용자와 컴퓨팅 서비스 기반 애플리케이션을 이어주는 인터페이스 역활을 한다. 온톨로지를 이용한 검색시스템은 빅 데이터 플랫폼과 SaaS를 같이 사용하면 시너지 효과를 낼 수 있다. 웹에서 서비스하고 있는 온톨로지를 수집하여 하나의 통합된 대규모 온톨로지를 구성하고 SaaS를 통한 빅 데이터 분석 솔루션을 이용하면 비용 효과적으로 실시간 시맨틱 검색 서비스를 제공할 수 있다.

웹 기반의 콘텐츠관리시스템은 클라우드에 콘텐츠관리 애플리케이션이 존재하고 웹브라우저를 사용하여 이를 이용하는 SaaS의 한 형태이다. 제안 모델의 응용을 위해서 [그림 2]와 같은 사설 클라우드 서버를 구축하였다. 사용자가 콘텐츠관리 애플리케이션에 접속을 요청하면 가상머신이 생성되고 그 위에서 동작하는 웹 서버의 콘텐츠 애플리케이션이 구동되어 서비스를 제공 받는 형태이다.

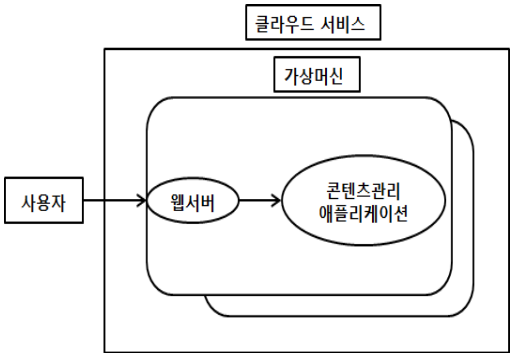

그림 2. 클라우드 컴퓨팅 모델

2.2 시맨틱 웹 모델

시맨틱 웹은 웹 문서를 상호 의미적으로 연결하여 관련된 웹문서를 쉽고 빠르게 찾을 수 있도록 온톨로지 형태로 표현하고 분산된 문서를 의미적 상호 운용성으로 통합하여 거대한 지식 베이스를 구축 하는 프레임워크이자 기술이다. 웹에 산재한 정보의 의미를 사용자가 파악하는 것이 아닌, 자동화된 기계가 해석할 수 있는 일종의 표준 의미정보 교환의 수단이 되는 온톨로지 형태로 표현되는 것이 시맨틱 웹의 목적이다. 시맨틱 웹의 기술은 명시적 메타데이터와 온톨로지 그리고 논리적 추론 등의 기술로 구성되어 있다. 명시적 메타데이터의 표현기술인 RDF는 XML에 기반한 시맨틱 마크업 언어로 <Subject, Predicate, Object>의 트리플 형태로 개념을 표현하는 반면 온톨로지는 이러한 트리플 구조에 기반하여 그래프 형태로 의미정보를 표현한다. 논리적 추론은 온톨로지와 함께 결합된 관계 정보들로부터 새로운 정보를 도출해 내는 기술이다. 온톨로지는 시맨틱 즉 내재된 의미를 공유하는 주요한 방법론의 하나로, 정보공유 뿐만 아니라 사용자와 컴퓨터간의 자유로운 통신을 가능케하고, 관련 지식을 표현, 저장할 수 있어 시스템의 활용 폭을 확장시킬 수 있다. 온톨로지는 도메인의 지식을 표현, 정리하기 때문에 도메인 전문가 등의 정보해석이 필요하다. 이는 사용자가 정보의 시맨틱을 쉽게 파악할 수 없기 때문에 별도의 전문가의 정보해석을 필요로 한다. 특정 분야에서만 사용되는 온톨로지와 달리 웹에서 사용되는 온톨로지는 지식 및 정보 표현의 범위가 한정되지 않다는 문제가 발생한다. 따라서 웹을 위한 온톨로지를 구축하기 위해서는 현재의 웹에 존재하는 데이터를 기준으로 정보에 의미를 부여하고 관계를 정의하면서 확장시켜 나가는 검색 키워드의 확장을 이용해 온톨로지를 생성하는 방법을 사용한다 [11-14].

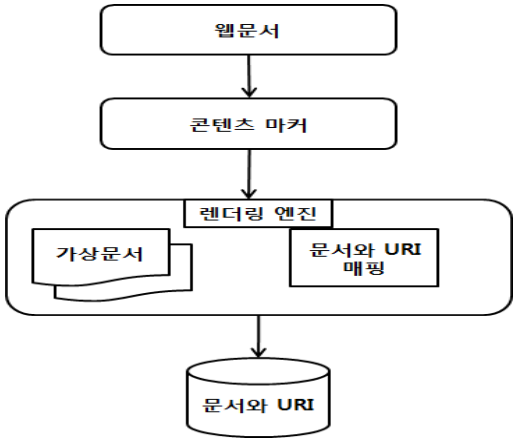

그림 3. 시맨틱 웹 모델

제안 모델의 응용을 위해서 [그림 3]와 같은 시맨틱 웹문서 관리시스템을 구축하였다. 검색 콘텐츠 마커는 검색 키워드의 확장을 통해 생성되고 데이터베이스에 저장된 온톨로지와 매핑을 통하여 웹문서에 시맨틱 어노테이션을 생성한다. 웹 문서 형태로 되어있는 비구조적 데이터를 시맨틱 어노테이션을 통해 데이터베이스에 저장된 온톨로지와 매핑하여 구조화된 정보를 얻어낼 수 있다. 렌더링 엔진은 데이터베이스에 저장되어 있는 ID를 갖는 도큐먼트, 대응되는 URL, 원본 콘텐츠를 참조하여 가상문서를 생성한다.

Ⅲ. 하둡기반 클러스터링 모델

3.1 하둡 분산시스템

웹 문서의 크롤링 양이 증가하면 비구조적 대규모 문서를 저장하기 위한 리포지터리로서 RDBMS는 확장에 따른 비용 증가와 처리지연이 발생하므로 비효율적이다. 이러한 문제의 해결책은 하둡 클라우드와 쉽고 빠르게 확장가능한 NoSQL 데이터베이스이다[22][23].

하둡과 NoSQL을 검색에 적용한 사례로는 페이스 북과 네이버의 검색 서비스가 있다. 이들 서비스 시스템은 실시간으로 정보를 색인하고, 색인이 됨과 동시에 검색이 가능하도록 하둡과 NoSQL을 활용하고 있다. 하둡의 맵 리듀스 모델과 NoSQL은 분산처리를 기반으로 하고 있는 빅 데이터 처리 기술로 검색에 소요되는 처리시간을 줄이고 데이터베이스의 수평적 확장을 용이하게 함으로써 구축비용을 감소시킨다.

제안 모델은 사설 클라우드 환경에서 SaaS로서 구현하고 리포지터리에는 URL, 인덱스, 타임스탬프, 콘텐츠를 저장하고 콘텐츠와 링크가 함께 연결된 구조를 갖도록 하였다.

시맨틱 웹은 RDF 기반으로 웹을 데이터 단위로 구조화 시켜 상호 관계성을 파악하려는 시도로 웹을 데이터베이스화 혹은 지식 기반으로 만들고 이를 기반으로 추론을 목표로 한다. 웹 문서에 포함된 콘텐츠의 관계성을 RDF로 추출한 후 URI를 기반으로 그래프를 따라 의미를 쫓아가는 방식이다. 하둡은 구글의 검색 데이터 처리에서 주로 사용한 맵 리듀스 방식을 오픈 소스 소프트웨어로 구현한 것으로 분산 환경의 데이터 처리와 저장을 쉽게 처리할 수 있다. [그림 4]와 같이 맵 리듀스 분산 환경에서 문서를 청크 단위로 쪼갠 후, 사용자 정의 함수인 매퍼에서 청크를 처리하고 리듀스 함수에서 원하는 결과값을 계산시켜 결과를 얻어낸다. 하둡을 검색 데이터 처리 기술에 사용하면 사회적 이슈가 발생했을 때, 분산 환경의 동적 제어 API를 이용하여 크롤링 및 인덱싱 작업을 비주기적으로 시행 가능하고, 비정형의 대용량 데이터 저장을 위한 NoSQL 플랫폼을 필요시 이용할 수 있어서 실시간 웹 검색을 위한 엔진 시스템으로서 적합하다[15-17]. [그림 5]는 웹 문서에 대한 URL과 콘텐츠의 저장패턴을 보여주고 있다.

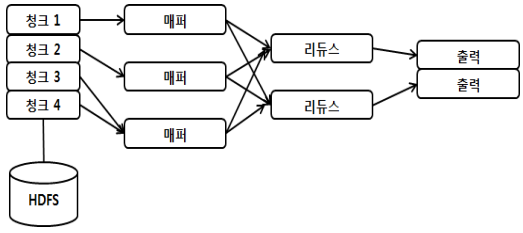

그림 4. 하둡 맵리듀스 분산 모델

그림 5. 웹 문서에 대한 저장 패턴

3.2 클러스터링 모델

디지털 정보의 거대한 성장으로 인해서 자동화된 문서 클러스터링은 콘텐츠관리시스템에서 필수적인 요소이다. 클러스터링 과정은 데이터 오브젝트간의 유사도를 분석하고 그룹화를 수행한다. 그룹화된 오브젝트는 데이터 셋의 대규모 목록을 통해서 쉽게 브라우징되도록 사용된다. 문서 클러스터링은 문서간의 유사도를 기반으로 문서를 그룹으로 분류하고 각각의 그룹은 유사한 문서들로 구성된다[18][19]. 각 문서들은 비지도학습을 통하여 높거나 낮은 유사도를 갖는 문서 클러스터로서 카테고리화 된다. 지도학습은 라벨링되어 있는 훈련용 데이터로부터 하나의 함수를 유추하기 위한 기계학

습인 반면 비지도학습은 라벨링이 되어있지 않는 훈련용 데이터를 사용하므로 함수를 추론할 수 없고 예측이 아닌 데이터가 어떻게 구성되어 있는지 밝히는데 주로 사용하는 일종의 그룹핑 기계학습 알고리즘이다. 전통적인 문서 클러스터링 방법에서 사용되는 벡터공간 모델은 관리 가능한 차원이 제한되는 문제점을 안고 있다. 제안 모델에서 응용되는 클러스터링 모델은 K-Means 클러스터링 알고리즘으로 [그림 6]과 같다. K-Means 클러스터링 알고리즘은 주어진 데이터를 K개의 클러스터로 묶는 알고리즘으로 클러스터 중심점과 최소의 거리를 갖는 데이터를 클러스터로 묶는 과정을 반복하여 단일 클러스터로 그룹화 되면 처리를 종료한다.

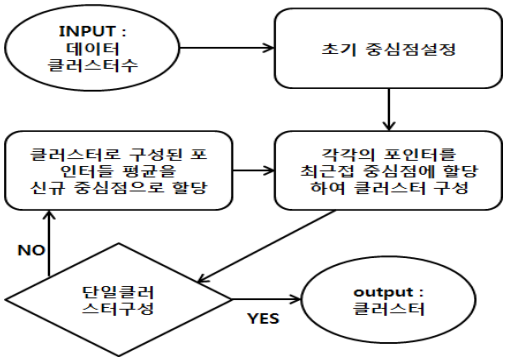

그림 6. K-Means 클러스터링 모델

Ⅳ. 키워드 빈도수와 컨셉 매칭 기반의 시맨틱 클러스터링 모델과 응용

4.1 문서와 검색어 기반 K-Means 클러스터링 모델

기존의 문서 검색 방식은 질의어가 포함된 문서 목록을 결과로 가져오며 사용자의 요구시에 내용을 제시하는 것이 일반적인 형태이다. 이는 웹 문서의 유사도와 시맨틱 관련성을 고려하지 않음으로써 사용자가 내용 검색과 분석에 많은 시간과 노력을 요구한다. 따라서 검색어를 키워드로 분리하고 이들 키워드와 문서 간의 유사도를 기반으로 [그림 7]과 같이 문서를 클러스터로 소속시켜서 분류하는 문서와 검색어 기반 K-Means 클러스터링 모델을 제안한다. 검색어의 키워드 빈도수를 가중치로 사용하여 문서를 분류하는 기법으로서 주어진 키워드 집합에 따라 문서를 특정 카테고리로 분류하고 문서를 관련된 내용 별로 자동으로 구조화함으로써 사용자가 많은 양의 문서들을 좀 더 편리하게 접근할 수 있게 해준다. 각각의 그룹들은 유사한 문서들로 구성되고 클러스터 내부의 유사도는 높다. 반면 다른 그룹 내에 있는 문서와는 서로 유사하지 않고 클러스터 사이의 유사도는 낮다. 문서와 검색어 기반 K-Means 클러스터링은 검색어를 키워드로 분리하고 키워드의 개수만큼 클러스터 개수를 설정한 후에 cluster[i][j]= argmax(freq(문서i,키워드j))에 대한 연산을 수행한다. 이 수식은 빈도수를 계산하는 freq함수를 최대화 시키는 색인 i와 j를 구하는 연산 수식으로 각각의 문서i를 모든 키워드j와 비교하여 출현 빈도가 가장 높은 키워드j 의 j와 i를 첨자값으로 갖는 cluster배열에 1의 값을 설정하여 문서의 그룹화를 표현한다. 클러스터링 완료 후 문서가 소속된 키워드 클러스터를 식별하기 위해서는 클러스터의 이차원 배열 값이 1인 경우의 i와 j를 출력하면 된다.

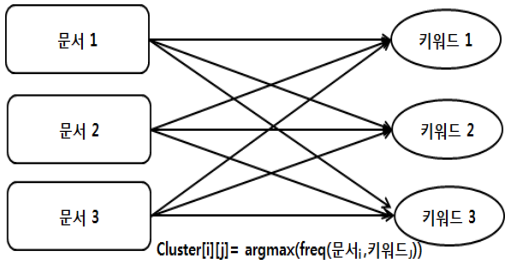

그림 7. 문서와 검색어 기반 K-means 클러스터링 모델

문서와 검색어 기반 K-means 클러스터링 모델에 대한 알고리즘은 알고리즘 1과 같다.

알고리즘 1.
클러스터 배열 초기화
for i=1 to n(문서의 개수)
for j=1 to n(검색어를 구성하는 키워드 개수)
  cluster[i][j]=0;

문서 클러스터 구성
for i=1 to n
cluster[i][j] =argmax(j = 1 to n) {freq(S[i], c[j]) }

4.2 컨셉 매칭을 통한 클러스터링 모델

지금까지의 검색엔진은 단순히 공통된 키워드를 기반으로 하여 검색된 문서를 사용자에게 제공하였다. 텍스트 마이닝 알고리즘이 X라는 문서와 Y라는 문서가 관련되고 Y라는 문서와 Z라는 문서가 관련되었다는 것을 분석해서 X라는 문서와 Z라는 문서의 관련성을 검색 결과로 제공한다면 사용자가 내용 검색과 분석에 드는 노력과 시간을 경감시킬 수 있다[20][21]. 따라서 검색 키워드의 확장을 통해 생성되어 데이터베이스에 저장된 온톨로지를 기반으로 [그림 8]과 같이 문서를 클러스터로 분류하는 컨셉 매칭을 통한 시맨틱 컴퓨팅 모델을 제안한다. 유사도를 계산하려는 두개의 문서를 스트링 배열로 분리하고 온톨로지의 컨셉 매핑을 통한 경로 정보를 식별한다. 식별된 경로 정보를 이용하여 두 문서 X와 Y의 유사도를 계산한다.

$SIM_{WP}(X,Y)=2*N/(N1+N2)$의 수식표현에서 N은 온톨로지에서 두 X와 Y 문서에 포함된 키워드 컨셉이 상호 연결되기 위해서 경유하는 최상위경로상의 컨셉과 온톨로지 루트 컨셉과의 아크 개수를 의미하고 N1과 N2는 X와 Y문서에 포함된 키워드 컨셉에서 온톨로지 루트 컨셉과의 아크의 개수를 의미한다.

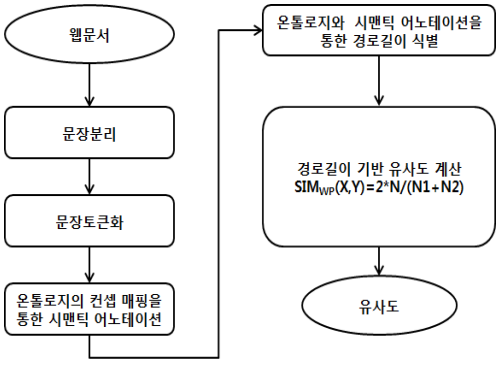

그림 8. 컨셉 매칭을 통한 클러스터링 모델

컨셉 매칭을 통한 클러스터링 모델에 의한 유사도 평가 알고리즘은 알고리즘2와 같고 문서 X와 문서 Y의 의미적 유사성 평가를 위해서 사용하는 R[i,j]는 X와 Y 문서에 포함된 각각의 키워드 컨셉에서 온톨로지 루트 컨셉과의 아크의 개수이고 |X|와 |Y|는 문서 X와 문서 Y의 워드의 개수를 의미한다.

알고리즘 2.

```
tot_X = 0;
tot_Y = 0;
int i=0;
while(i < |X|) {
  max_i=0;
  int j=0;
  while(j < |Y|) {
    if (R[i, j] > max_i) max_i=R[i, j] > max_i;
    j++
  }
  tot_X += max_i;
  i++;
}
int j=0;
while(j < |Y|){
  max_j=0;
  int i=0;
  while(i < |X|){
    if (R[i, j] > max_j)
        max_j=R[i, j] > max_j;
    i++;
  }
    tot_Y += max_j;
    j++;
}
similarity = (tot_X + tot_Y) / 2 * (|X| + |Y|)
```

4.3 매핑을 통한 클러스터링 결과 분석

키워드 빈도수와 컨셉 K-Means 클러스터링 모델은 n개의 데이터를 k개의 그룹으로 군집화 하는 모델로 데이터 간의 유사성(similarity)은 거리(Distance)를 가지고 측정하고 클러스터의 중심점과 최소의 거리를 갖는 데이터를 클러스터로 그룹화 하는 반면 키워드 빈도수를 통한 클러스터링은 K-Means 클러스터링 모델을 변형하여 키워드의 개수를 클러스터의 수로 설정하고 키

워드가 포함되어 있는 문서들을 묶어서 그룹화하고 문서간의 유사성은 키워드의 빈도수를 가중치로 사용하여 계산한다. 컨셉 매칭을 통한 클러스터링은 문서의 스트링을 온톨로지의 컴셉과의 매칭을 통해서 식별된 경로 정보를 사용하여 관련성 점수를 평가하여 유사도를 계산하여 군집화를 수행한다. 문서와 검색어 기반의 클러스터링에 소요된 시간은 [그림 9]와 같고 컨셉 매칭을 통한 클러스터링에 소요된 시간은 [그림 10]과 같다. 클러스터링에 걸린 시간은 Y축에 클러스터는 X축에 표시하였다. 50개의 문서가 테스트를 위해서 사용되었고 3개의 키워드가 검색어로 사용하였다.

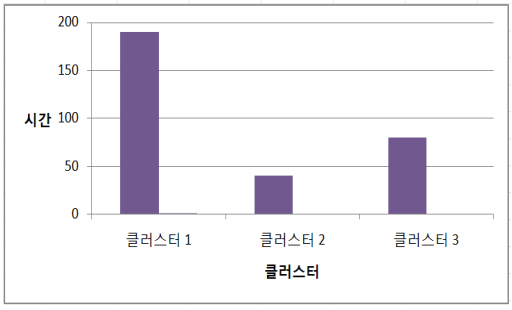

그림 9. 문서와 검색어 기반의 클러스터 구성

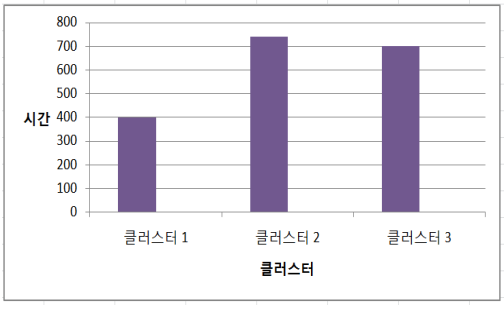

그림 10. 컨셉 매핑 기반의 클러스터 구성

컨셉 매칭을 통한 클러스터링이 문서와 검색어 기반의 클러스터링보다 웹문서를 분류하는데 오랜 시간이 걸린 다는 것을 확인하였다. 이는 컨셉 매칭을 통한 클러스터링이 웹 문서를 온톨로지의 컨셉과의 매핑을 통한 경로 정보를 획득하여 두 웹 문서의 유사도를 계산하기 때문이다.

V. 결론

기존의 웹 문서의 검색은 키워드를 이용하여 해당하는 질의어가 포함된 문서 목록을 결과로 가져오며 사용자의 요구시에 내용을 제시하는 것이 일반적인 형태이다. 이는 웹 문서의 유사도와 시맨틱 관련성 고려하지 않음으로써 사용자가 내용 검색과 분석에 많은 시간과 노력을 요구한다. 이의 해결을 위해서 빅데이터 요소 기술인 하둡과 NoSQL을 활용하여 대용량 시맨틱 웹 문서에 포함된 질의어의 키워드 빈도에 기반한 웹 문서의 유형 분류와 유사도를 제시하는 시맨틱 컴퓨팅 모델을 제안하고 구현하였다. 클라우드 환경에서 문서의 유형 분류를 위한 시맨틱 클러스터링 모델로 제안한 아키텍처와 알고리즘은 블로그나 소셜 네트워크에서 웹 콘텐츠의 폭증에 따른 검색 문제를 해결하기 위한 비즈니스 모델로 응용되어 사용될 수 있고 실시간 데이터 처리가 요청되는 이종 모델을 가진 공공 데이터와 웹 데이터를 취합하여 일반 사용자가 쉽게 질의할 수 있는 대용량 지식 기반 시스템을 구축하는데 응용 모델로 활용 될 수 있다. 본 연구의 한계점으로는 온톨로지를 바탕으로 검색어와 유사한 키워드를 찾고 이들 검색어와 유사 키워드가 문서에서 출현하는 빈도를 이용하여 시맨틱 클러스터링을 수행하는 모델을 제안하고 있으나 웹문서에 시맨틱 정보를 직접 입력하여 온토로지를 생성함으로써 온톨로지의 범위를 제한하고 있는 한계를 가지고 있다. 따라서 향후 웹 온톨로지를 자동으로 생성하는 시스템에 대한 연구를 지속할 예정이다.

참 고 문 헌

[1] 김영수, 문형진, 조혜선, 김병익, 이진해, 이진우, 이병엽, “계층적침해자원기반의 침해사고 구성 및 유형 분석,” 한국콘텐츠학회논문지, 제16권, 제11호, pp.139-153, 2016.

[2] 김영수, “보안 인텔리전트 유형 분류를 위한 다중 프로파일링 앙상블 모델,” 한국콘텐츠학회논문지, Vol.17, No.3, pp.231-237, 2017.

[3] 이태휘, 임동혁, "맵리듀스에서의 구조적 RDF 데이터 변경 탐지 기법," 정보처리학회논문지, Vol.3, No.8, pp.293-298, 2014,

[4] 심준, 이홍철, "검색 키워드 확장을 이용한 온톨로지 자동 생성 시스템 개발," 한국산학기술학회논문지, Vol.10, No.6, pp.1220-1228, 2009.

[5] 배우정, 이현영, 박인철, 이용석, "개념 그래프의 트리 표현," 한국정보과학회 학술발표논문집, Vol.25(1B), pp.393-395, 1998.

[6] 안윤선, 김윤희, "군집분석을 이용한 하이브리드 클라우드 컴퓨팅 환경에서의 시맨틱 클라우드 자원 추천 서비스 기법," 정보처리학회논문지, Vol.,4 No.9, pp.283-288, 2015.

[7] P. Mell and T. Grance, "The NIST definition of cloud computing," National Institute of Standards and Tchnology, Vol.53, No.6, p.50, 2009.

[8] C. N. Hoefer and G. Karagiannis, Taxonomy of cloud computing services. In: GLOBECOM Workshops (GC Wkshps), 2010 IEEE. pp.1345-1350, IEEE 2010.

[9] P. Bhaskar, J. Admela, K. Dimitrios, and G. Yves, "Architectural Requirements for Cloud Computing Systems:An Enterprise Cloud Approach," J. Grid Computing, Vol.9, No.1, pp.3-26, 2011.

[10] Wei-Tek Tsai, Xin Sun, and Janaka Balasooriya, "Service-Oriented Cloud Computing Architecture," 2010 Seventh International Conference on Information Technology, 2010.

[11] H. Rijgersberg, M. Wigham, and J. T. Top, "How semantics can improve engineering processes: A case of unitsof measure and quantities," Advanced Engineering Informatics, Vol.25, No.2, pp.276-287, 2011.

[12] P. Shvaiko and J. Euzenat, Ontology Matching: State art and Future Challenges, pp.1-15, IEEE 2013.

[13] K. Saruladha, G. Aghila, and B. A. Sathiya, "Comparative Analysis of Ontology and Schema Matching Systems," International Journal of Computer Application, Vol.34, No.8, pp.14-21, 2011.

[14] A. Ismail and M. Joy, Semantic searches for extracting similarities in a content management system. Proceedings the IEEE International Conference on Semantic Technology and Information Retrieval, Putrajaya, pp.113-118, June 28-29, 2011,

[15] N. Leavitt, Will NoSQL databases live up to their promises. computer, Vol.43, pp.12-14, 2010.

[16] R. P. Padliy, M. R. Patra, and S. C. Satapthy, RDBMS to NoSQL: Reviewing some next-generation non-relational databse's. Int J. Adv. Eng. Sci Techno1, Vol.11, pp.15-30, 2011.

[17] David Sánchez, Montserrat Batet, David Isern, and Aida Valls, "Ontology-based semantic similarity: A new feature-based approach," Journal of Expert systems with applications, Elseveir, No.39, pp.7718-7728, 2012.

[18] J. H. Hwang and K. H. Ryu, "A weighted common structure based clustering technique for XML documents," Elsevier Publication, 2010.

[19] B. Drakshayani and E V Prasad, "Text Document Clustering based on Semantics," International Journal of Computer Applications, pp.0975-8887, Vol.45, No.4, May 2012.

[20] R. Priyadarshini and Latha Tamilselvan, "Document clustering based on keyword frequency and concept matching technique in Hadoop," International Journal of Scientific & Engineering Research, Vol.5, Issue 5, May, 2014.

[21] R. Priyadarshini, Latha Tamilselvan, "Document Based Semantic CMS in Cloud," Information Technology Journal, Vol.13, pp.217-230, February 07, 2014.

저 자 소 개

김 영 수(Young Soo Kim) 정회원

- 2003년 8월 : 국민대학교정보관리학(정보관리학박사)
- 현재 : 충남 재활IT 융합 기술원 대표 컨설턴트
- 현재 : 배재대학교 사이버보안학과

<관심분야> : 빅데이터서비스보안, 정보 보안

이 병 엽(Byoung Yup Lee) 종신회원

- 1991년 2월 : 한국과학기술원 전산학과(공학사)
- 1993년 2월 : 한국과학기술원 전산학과(공학석사)
- 1997년 2월 : 한국과학기술원 경영정보공학(공학박사)
- 1993년 1월 ~ 2003년 2월 : 대우정보시스템 차장
- 2003년 3월 ~ 현재 : 배재대학교 사이버보안학과 교수

<관심분야> : XML, 지능정보시스템, 데이터베이스 시스템, 전자상거래학