JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 18, No. 9 : 2155~2160 Sep. 2014

하둡과 의미특징을 이용한 문서요약

김철원*

Document Summarization using Semantic Feature and Hadoop

Chul-Won Kim*

Department of Computer Engineering, Honam University, Gwangju 506-714, Korea

요 약

본 논문은 하둡 기반의 분산병렬처리에 의한 문서의 의미특징을 추출하고, 추출된 의미특징을 이용하여 문서를 요약하는 새로운 방법을 제안한다. 제안된 방법은 문서요약에 비음수 분해된 문서의 의미특징을 이용함으로써 문서의 내부 구조를 잘 표현 할 수 있다. 또한 하둡을 이용하여 빅데이터의 문서를 요약할 수 있다. 실험결과 제안방법이 단일 컴퓨터 환경에서 처리할 수 없는 대용량의 문서를 요약할 수 있음을 보인다.

ABSTRACT

In this paper, we proposes a new document summarization method using the extracted semantic feature which the semantic feature is extracted by distributed parallel processing based Hadoop. The proposed method can well represent the inherent structure of documents using the semantic feature by the non-negative matrix factorization (NMF). In addition, it can summarize the big data document using Hadoop. The experimental results demonstrate that the proposed method can summarize the big data document which a single computer can not summarize those.

키워드 : 문서요약, 의미특징, 하둡, 분산병렬처리

Key word : Document summarization, semantic feature, Hadoop, distributed parallel processing

접수일자 : 2013. 11. 04 심사완료일자 : 2013. 12. 04 게재확정일자 : 2013. 12. 20

* **Corresponding Author** Chul-Won Kim (E-mail:cwkim@honam.am.kr,Tel:+82-62-940-5403)
Department of Computer Engineering,Honam University,Gwangju 506-714,Korea

Open Access http://dx.doi.org/10.6109/jkiice.2014.18.9.2155 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

통신기기의 발전은 사용자들로 하여금 인터넷 상의 정보를 더욱 쉽게 접근할 수 있게 하고 있으며, 이로 인하여 인터넷 상의 정보 증가를 더욱 가속화 시키고 있다. 또한 인터넷 사용이 가능한 소형 통신기기들의 보급 증가는 대용량의 인터넷 정보를 소형 통신기기의 화면에 효율적으로 표시할 수 있는 정보 요약 기술에 대한 필요성을 증가시키기고 있다.

인터넷 정보의 요약 기술은 검색엔진의 요약문인 스니핏, 정보추출, 텍스트마이닝, 질의응답 분야 등에서 다양한 통계적 방법론을 사용하여 요약의 정확률을 향상시키기 위해 연구되어 왔다. 그러나 기존 연구는 단일 코어 또는 단일 컴퓨터 상황에서 수행되기 때문에, 현재와 같인 폭발적으로 증가하는 인터넷 및 소셜 네트워크 서비스 등의 대량의 정보에 대한 신속한 요약 요구를 처리할 수 없는 상황에 직면하고 있다. 즉, 지금까지 정보 요약에 관한 연구는 다양한 통계적 방법론을 이용하여 요약정보의 성능향상에 목적을 두고 수행되어 높은 수준의 요약율을 보여주고 있다. 그렇지만 대부분의 문서요약 연구가 정확률에 맞추어져 있어 대용량의 자료에 대한 요약속도 또는 요약의 처리율에 대한 연구를 미흡한 실정이다.

본 연구는 대용량의 문서로부터 정보를 요약할 수 있는 문서요약방법을 연구한다. 제안방법은 대용량의 문서를 전처리하여 하둡[1]에 분산저장하고, 분산저장된 자료로부터 문서의 의미특징을 병렬처리하여 추출하며, 추출된 의미특징을 이용하여 문서를 요약한다. 제안방법은 PC(personal computer, 개인용 컴퓨터)로 구성된 하둡 기반의 노드에 대량의 문서 자료를 저장함으로써 단일 PC에 처리할 수 없는 대용량의 문서를 요약할 수 있는 장점을 갖는다. 하둡은 분산 파일 시스템(distribution file system)과 분산 컴퓨팅을 위한 맵리듀스(MapReduce)를 포함하여 개발된 분산병렬처리 시스템이다[1].

본 논문의 구성은 다음과 같다. 2장에서는 관련연구로 문서요약과 하둡 프레임워크에 대하여 알아본다. 3장에서는 제안방법인 문서요약 방법을, 제4장에서는 실험 및 평가에 대해 기술한다. 마지막으로 제5장에서는 결론을 맺는다.

Ⅱ. 관련연구

2.1. 문서요약

다음은 본 논문의 제안방법과 관련된 최근의 문서요약 연구들이다. Nastase는 위키피디아와 워드넷의 외부지식과 확장 활동에 의한 주제기반 다중문서요약 방법을 제안하였다. 이들의 방법은 외부지식을 이용하여 사용자의 질의를 확장하였으며, 확장된 질의와 관련된 문서를 문법적으로 연결된 용어의 그래프로 표현하여 문서를 요약하였다[2]. Ramanathan외 저자들은 언어에 독립적인 단일문서요약 방법을 제안하였다. 이들의 방법은 문서의 문장과 위키피디아의 의미적 개념과 일치시켜 문서를 요약한다[3]. Ye와 저자들은 위키피디아 기반으로 새로운 문서 개념격자를 이용한 문서요약 방법을 제안하였다. 이들이 제안한 방법은 문장 간의 연관관계를 얻기 위하여 위키피디아의 내부 연결에 의한 위키피디아 개념을 소개하였으며, 위키 개념과 비 텍스트 특징을 이용하여 확장된 문서 개념 격자 모델(extended document concept lattice model)을 제안하였다[4]. Gong외 저자들은 위키피디아를 이용한 요약 시스템을 제안하였다. 이들이 제안한 방법은 위키피디아의 개념을 인식한 후에, 개념에 가중치를 주고 특징을 개산하여 요약문을 생성한다[5]. 위키피디아를 이용한 문서요약 방법은 문서요약 하기 전에 대량의 위피키디아 자료를 전처리하거나 학습함으로써 많은 자원을 소모해야 하는 문제를 가지고 있다.

Sanderson은 문서상의 중요한 문장과 사용자가 확장한 질의를 이용하여 문서를 요약 방법을 제안하였다[6]. Tombros와 Sanderson은 문서의 형식에 포함된 정보인 제목, 주제, 용어의 빈도 정보, 질의 등을 점수화 하여서 사용자가 보조 정보로 활용할 수 있는 문서 요약 방법을 제안하였다[7]. Varadarajan과 Hristidis는 질의와 가장 관련이 높은 문장과 의미 연관을 이용하여서 문서로부터 추출된 복합 주제를 적용하여서 질의에 특화된 문서 요약 방법을 제안하였다[8].

2.2. 하둡 프레임워크

하둡 프레임워크(hadoop framework)는 응용프로그램들에게 대용량의 데이터를 처리하게 할 수 있도록 지원하는 분산 프레임워크이다. 하둡은 아파치 프로젝트의 하나로 분산파일시스템(HDFS: hadoop distributed

file system)과 맵리듀스(MapReduce)를 포함하고 있다. HDFS는 낮은 성능의 컴퓨터에도 대용량의 데이터를 처리할 수 있도록 지원하는 파일 시스템으로 고가용성(fault-tolerant)을 제공하고 있다[1].

HDFS는 하나의 마스터 노드와 여러 개의 슬레이브 노드로 구성된다. 마스터 노드는 파일 시스템 네임스페이스를 관리하고 클라이언트에 의한 파일 접근을 통제하는 단일 네임 노드로 구성된다. 각 슬레이브 노드는 저장장치를 관리하는 데이터 노드가 있다 . HDFS는 파일을 하나 또는 그 이상의 블록으로 분할하여 데이터 노드에 저장한다. 네임 노드는 파일과 디렉토리의 열기, 닫기, 이름 변경과 같은 파일 시스템의 네임스페이스 동작을 수행하고 데이터 노드의 블록 일치 검사하여 판단한다. 데이터 노드는 네임 노드의 지시로 블록 생성, 삭제, 복제를 수행한다[1].

맵리듀스는 분산 및 병렬로 대용량의 자료를 효율적으로 관리하기 위하여 HDFS를 기반으로 개발되었다. 맵리듀스는 병렬처리, 고가용성, 데이터 분산 및 로드 밸런싱 등을 지원한다. 맵리듀스는 맵(map)과 리듀스(reduce)의 합성어로 맵 함수와 리듀스 함수의 조합으로 분산 병렬 응용프로그램을 지원한다. 맵리듀스는 데이터를 {key, value}의 쌍으로 만들어 분산 병렬처리가 가능하도록 지원한다. 맵은 사용자 정의 자료구조이며, 입력 데이터에서 {key, value} 쌍으로 구성된 중간 데이터를 만들어 리듀스에 전달한다. 리듀스 함수는 key 값을 이용하여 value 값을 합쳐 출력한다[1].

Ⅲ. 제안방법

본 논문에서는 분산병렬처리 기반의 문서요약 방법을 제안한다. 제안방법은 전처리, 분산병렬 의미특징 추출, 문서요약 단계로 구성된다. 전처리 단계에서는 문서를 하둡에 용어문장빈도행렬을 분산저장하고, 분산병렬 의미특징 추출단계에서는 분산저장된 용어문장 빈도행렬을 병렬로 비음수행렬하여 의미특징을 추출한다. 마지막 요약 단계에서는 계산된 의미특징을 이용하여 문서를 요약한다. 다음 그림1은 제안된 하둡 기반의 문서요약방법의 개요이다.

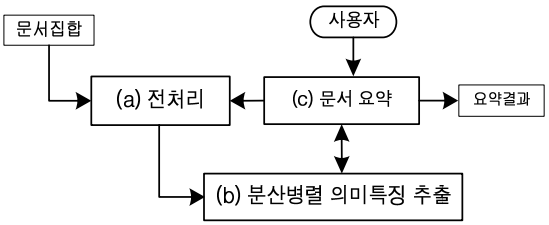

그림 1. 하둡 기반의 문서요약 방법정보
Fig. 1 Document summarization method based on Hadoop

3.1. 전처리

일반적으로 문서요약을 위한 전처리는 문서집합에 불용어를 제거하고, 용어의 어근을 추출하여 용어문장 빈도행렬을 생성한다. 그러나 본 논문의 전처리 단계는 문서요약에 분산병렬처리가 가능하도록 고려해야 한다. 이를 위해서 본 논문에서는 다음과 같이 문서집합을 전처리하여 용어문장 빈도행렬을 하둡에 분산저장한다. 첫째, 문서집합을 문장집합으로 분해한다. 둘째, 문장집합을 머하우의 문서 벡터를 만드는 도구를 이용하여 용어문장 빈도행렬을 만든다. 즉, 머하우(Mahout)[9]도구인 SequenceFileFromDirectory 클래스의 seqdirectory와 SparseVectorFromSequenceFiles 클래스의 seq2sparse를 이용한다. SequenceFileFromDirectory는 디렉토리에 속한 문장집합을 SequenceFile 형식으로 표현한 중간 단계의 문서를 생성한다. SparseVectorFromSequenceFiles는 SequenceFile 형식의 문서를 용어문장 빈도행렬로 변환한다. 다음 표1은 두 가지 도구를 이용하여 용어문장 빈도행렬을 구성하는 예제를 보여준다.

표 1. 머하우를 이용한 용어문장 빈도행렬 생성
Table. 1 Construction of term-sentence frequency matrix using mahout

```
bin/mahout seqdirectory -c UTF-8 -i summary/term/ -o
term-seqfiles
// summary 디렉토리의 문장집합인 term 파일일
   중간파일인 term-seqfiles로 변호나

bin/mahout seq2sparse -i term-seqfile/ -o term-vectors -ow
// term-seqfiles 파일을 용어문장 빈도행렬인 term-vectors
   파일로 변환
```

표1의 도구를 이용하여 용어문장 빈도행렬을 생성하면 하둡에는 df, tf, tfidf, dictionary, tokenized-sentences 등의 값이 각각의 디렉토리에 분산저장된다. 여기서 저장되는 값들은 다음과 같다. df에는 문장의 빈도수, tf에는 용어의 빈도수, tfidf는 tf와 df를 이용한 가중치 벡터문서, dictionary에는 용어와 숫자로된 ID와 매핑 내용, tokenized-sentences는 문장을 개별단어로 쪼개어 저장된다.

3.2. 분산병렬 의미특징 추출

문서요약은 문서를 대표할 수 있는 중요한 특징을 나타내는 의미특징(semantic feature)을 이용함으로써 요약의 질을 높일 수 있다. 주로 이용되는 문서의 내부특징을 추출하는 방법으로는 비음수 행렬분해, 주성분분석 등이 있으며 비음수 행렬분해가 가장 의미 있는 특징을 추출하는 것으로 알려져 있다. 그러나 이들 방법은 문서 내부에 잠재되어 있는 고유한 특징(inherent feature)을 행렬분해하여 추출함으로써 문서의 양이 많아질수록 계산 량이 많아져 의미특징 추출이 제한되는 문제점을 가지고 있다.

본 논문은 문서양의 증가시 의미특징 추출이 제한되는 문제를 해결하기 위해 하둡을 이용하여 비음수 행렬분해가 분산병렬처리에 의해 계산이 가능하도록 제안한다. 비음수행렬분해 알고리즘은 다음과 같이 계산된다. 문장집합이 k개의 의미특징 벡터로 구성된다고 가정할 때, 행렬 A를 식(2)의 목적 함수가 최소 값을 갖도록 식(3)을 반복하여서 식(1)과 같이 비음수의미특징 행렬(NSFM, non-negative semantic feature matrix) W와 비음수의미변수행렬(NSFM, non-negative variable matrix) H로 분해한다[10].

$$A \approx WH \tag{1}$$

$$J = \frac{1}{2} \| A - WH \| \tag{2}$$

여기서 행렬 A는 $m \times n$개의 원소로 구성되는 행렬이며, 행렬 W 는 $m \times k$개의 원소로 구성되고, 행렬 H는 $k \times n$개의 원소로 구성된다.

$$W_{ij} \leftarrow W_{ij} \frac{(AH^T)_{ij}}{(WHH^T)_{ij}}, \; H_{ij} \leftarrow H_{ij} \frac{(W^TA)_{ij}}{(W^TWH)_{ij}} \tag{3}$$

여기서 H^T는 H의 전치행렬이고, W^T는 W의 전치행렬이다.

본 논문에서는 비음수행렬 분해를 하둡을 이용하여 분산병렬로 계산하기 위하여 Liu이가 제안한 맵리듀스(MapReduce)방법을 이용한다. Liu의 방법[11]은 식(3)을 3계의 단계로 분해하여 맵과 리듀스로 구성한다. 즉 다음 표2와 같이 맵과 리듀스로 분해하는 단계를 요약할 수 있다.

표 2. Liu의 맵리듀스를 이용한 비음수행렬분해 방법
Table. 2 Liu's NMF method using MapReduce

단계	$W_{ij} \leftarrow W_{ij} \frac{(AH^T)_{ij}}{(WHH^T)_{ij}}$	$H_{ij} \leftarrow H_{ij} \frac{(W^TA)_{ij}}{(W^TWH)_{ij}}$
1	$X1 = WH^T$를 Map과 Reduce로 분해 계산	$X2 = W^TA$를 Map과 Reduce로 분해 계산
2	$Y1 = WHH^T$를 Map과 Reduce로 분해 계산	$Y2 = W^TWH$를 Map과 Reduce로 분해 계산
3	$W = W\frac{X1}{Y1}$를 Map과 Reduce로 분해 계산	$H = H\frac{X2}{Y2}$를 Map과 Reduce로 분해 계산

3.3. 문서요약

비음수 행렬분해에 의해 추출된 의미특징 행렬의 의미는 다음과 같다. 의미특징행렬 W는 문장 집합에 포함된 용어들에 문장집합에 잠재되어 있는 특징을 값(value)로 표현한 것이며, 의미변수행렬 H는 문장집합에서 각각의 문장이 잠재되어 있는 특징을 값으로 표현된다. 또한 의미특징 행렬 W의 각각의 값들은 의미변수행렬 H의 각각 열벡터에 대한 가중치의 의미를 갖는다.

이러한 문장집합과 의미특징 행렬간의 특성을 이용하여 문서를 요약할 수 있다. 본 장에서는 사용자의 질의와 의미특징행렬 W와 코사인 유사도를 식(4)를 이용하여 계산하고 유사도가 가장 높은 의미특징 열벡터를 찾는다. 찾은 의미특징 열벡터에 일치하는 의미변수행렬 H의 행벡터에서 가장 높은 값을 갖는 의미특징을 선택한다. 선택된 의미특징과 일치하는 문장을 요약문으로 추출한다.

$$sim(W_{*j}, q) = \frac{W_{*j} \cdot q}{|W_{*j}| \times |q|} \quad (4)$$

여기서 W_{*j}는 의미특징행렬 W의 j번째 열벡터를 나타내고, q는 사용자의 질의를 나타낸다[12].

Ⅳ. 실험 및 평가

본 논문에서 제안한 방법의 성능 평가하기 위한 자료를 야후(www.yahoo.com)에서 수집하여 용량별로 만들었다. 평가방법은 용량별 수집 자료를 사용자의 500개 질의를 이용하여 500개의 요약문을 생성하는 시간을 단일 환경과 4개의 노드를 분산병렬 간에 비교하여 평가하였다. 다음 표 3은 실험에 사용된 컴퓨터 환경을 나타낸다.

표 3. 실험 환경
Table. 3 Experiment environment

구분	단일 환경	4 노드 환경
하드웨어	i3 CPU * 1 8 GB RAM 1 TB HDD	i3 CPU * 4 8 GB RAM * 4 1 TB HDD * 4
운영제체	CentOS 6.3	CentOS 6.3
기타	-	하둡

다음 표4는 실험결과를 나타낸다. 여기서 용어문장 빈도행렬의 비음수 행렬분해시 생성되는 의미특징 행렬의 크기를 결정하는 의미특징 k의 개수는 200으로 고정한다. 의미특징의 개수가 늘어날수록 계산 량이 기하급수로 늘어난다. 표4의 실험결과를 보면 단일환 경시 100MB이상의 자료는 거의 계산이 안 되는 것을 볼수 있다. 이것은 문서요약의 비음수행렬분해에 사용되는 메모리의 크기가 RAM의 크기를 초과하기 때문에 거의 계산이 불가능한 것을 알 수 있다. 그러나 4대의 컴퓨터를 하둡상에서 분산병렬처리에 의한 문서요약의 비음수행렬 계산시 저장자료의 크기에 따라서 시간은 증가하나 문서가 요약됨을 알 수 있다.

표 4. 실험 결과환경
Table. 4 Experiment results

구분	100MB	500MB	1GB	10GB	20GB
단일 환경	181분	×	×	×	×
4 노드 환경	21분	67분	123분	732분	1632분

Ⅴ. 결 론

본 논문에서는 분산병렬처리 기반의 문서요약 방법을 제안하였다. 제안방법은 전처리된 문서를 하둡에 용어문장빈도행렬을 분산저장하고, 분산병렬로 의미특징을 추출하여 사용자자의 질의와 가장 잘 부합되는 문장으로 요약한다. 본 논문은 다음과 같은 장점을 갖는다. 문서요약에 의미특징을 이용하여 사용자가 원하는 요약문장을 추출함으로써 요약문의 질을 높였다. 또한 4대의 PC기반의 하둡 클러스터를 이용하여 대용량의 자료를 분산병렬처리 함으로써 단일 환경에서 처리할 수 없는 대량의 문서들을 처리할 수 있게 했다. 본 논문의 연구결과는 검색엔진의 스니핏 추출, 정보추출, 인터넷 대용량 자료의 요약 분석, 소셜 네트워크 서비스 대용량 자료의 요약 분석, 소형 인터넷 단말기의 정보요약 제공 등 기타 대용량 자료를 요약하는 분야에 활용할 수 있다. 앞으로 추가 연구로는 현재 실험환경상 4대의 노드만을 이용하였으나, 추가 노드 환경을 구성하여 다양설정에 따른 최적화 연구가 필요하다.

감사의 글

이 논문은 2013년도 호남대학교 학술연구비 지원을 받아 연구되었음

REFERENCES

[1] T. White, *Hadoop: The Definitive Guide,* 3th ed. O'Reilly Media, 2012.

[2] V. Nastase, "Topic-Driven Multi-Document Summarization with Encyclopedic Knowledge and Spreading Activation," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* Honolulu, Hawaii, USA, pp.763-772, 2008.

[3] K. Ramanathan, Y. Sankarasubramaniam, N. Mathur, A. Gupta, "Document Summarization using Wikipedia", in *Proceedings of the First International Conference on HCI*, Japan, 2009.

[4] S. Ye, T. S. Chua, J. Lu, "Summarization Definition from Wikipedia", in *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, Singapore, pp. 199-207, 2009.

[5] S. Gong, Y. Qu, S. Tian, "Summarization using Wikipedia", in *Proceedings of Text Analysis Conference 2010*, Gaithersburg, Maryland, USA, 2010.

[6] M., Sanderson, "Accurate user directed summarization from existing tools", in *Proceeding of the international conference on information and knowledge management,* Bethesda, Maryland, USA, pp.45-51, 1998.

[7] A., Tombros, M., Sanderson, "Advantages of Query Biased summaries in Information Retrieval", in *Proceeding of ACM Special Interest Group on Information Retrieval*, pp.2-10, Melbourne, Australia, 1998.

[8] R., Varadarajan, V., Hristidis, "A System for Query Specific Document Summarization", *in Proceeding of the International Conference on Information and Knowledge Management*, Arlington, Virginia, USA, pp.622-631, 2006.

[9] S. Owen, R. Anil, T. Dunning, E. Friedman, *Mahout in Action*, Manning Publiications, 2011.

[10] D. D. Lee, H. S. Seung, "Algorithms for non-negative matrix factorization," I*n Advances in Neural Information Processing Systems*, vol. 13, pp.556-562, Aug. 2001.

[11] C. Liu, H. C. Yang, J. Fan, L. W. He, Y. M. Wang, "Distributed Nonnegative Matrix Factorization for Web-Scale Dyadic Data Analysis on MapReduce," in *Proceeding of the International World Wide Web Conferene Comittee*, USA, pp.1-10, 2010.

[12] B. Y. Ricardo, Berthier, R. N., *Moden Information Retrieval*, ACM Press. 1999.

김철원(Chul-won Kim)

1997년 광운대학교 (공학박사)
1988년 ~ 현재 호남대학교 컴퓨터공학과 교수
※관심분야 : XML 응용, 멀티미디어 정보검색, 멀티미디어 정보처리 및 응용