클라우드 기반의 가중치에 의한 문서요약

박선* · 김철원*

*국립목포대학교, **호남대학교

Document Summarization using Weighting based on Cloud

Sun Park* · Chul Won Kim**

*Mokpo National orea Maritime University, **Honam University

E-mail : sunpark@mokpo.ac.kr, cwkim@honam.ac.kr

요 약

본 논문은 클라우드 기반의 가중치에 의한 문서요약 방법을 제안한다. 제안된 방법은 연관피드백을 이용하여 사용자의 간섭을 최소화 시키며, 클라우드 기반의 비음수 행렬분해를 이용한 의미특징으로부터 유도된 용어의 가중치는 문장집합의 내부 특징을 잘 나타나기 때문에 문서요약의 질을 향상할 수 있다.

ABSTRACT

In this paper, we proposes a document summarization method using the weighting based on cloud. The proposed method can minimize the user intervention to use the relevance feedback. It also can improve the quality of document summaries because the inherent semantic of the sentence set are well reflected by term weighting derived from semantic feature using nonnegative matrix factorizaitno based cloud.

키워드

문서 요약(document summarization), 의미특징(sematic features), 용어 가중치(term weighting), 비음수 행렬분해(NMF), 클라우드(cloud)

Ⅰ. 서 론

문서요약에 대한 접근방법은 통계적 방법, 그래프기반 방법, 언어학기반 방법, 의미정보기반 방법, 외부자원기반 방법, 기타 복합기반 방법이 있다[1-11].

본 논문은 문서요약 과정에서 병목현상이 발생하는 작업을 후보요약문장 추출과정의 언어처리 부분과 연관피드백을 이용하여 용어의 가중치 할당 과정의 통계적으로 취합하는 부분으로 분류하고, 각 작업을 클라우드 기반의 하둡의 맵 작업과 리듀스 작업을 이용하여 해결하는 분산병렬처리 기반의 문서요약 방법에 대하여 연구한다.

Ⅱ. 본 론

본 논문에서는 분산병렬처리 기반의 문서요약 방법에 대하여 연구 한다. 본 논문의 내용은 그림 1 문서 요약 모듈과 그림2 분산병렬처리 모듈로 구성된다. 문서요약 모듈에서는 그림(a) 전처리와 그림(b) 요약 알고리즘을 연구하며, 분산병렬처리 모듈에서는 PC 3대로 구성된 그림(d) 하둡 클러스터를 이용하여 분산저장하며, 그림(c) 하이브를 이용한 요약의 병렬처리에 대한 연구를 진행한다.

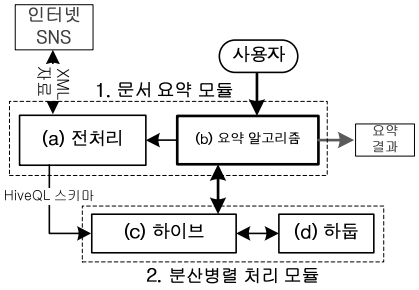

그림1. 제안방법

III. 결 론

본 논문에서는 하둡기반의 클라우드와 의미특징기반의 용어 가중치에 의한 문서요약 방법을 연구한다. 본 연구의 기여되는 다음과 같다. 첫째, 연구된 방법은 의사연관피드백을 이용하여 사용자의 간섭을 최소화 시키며, 의미특징으로부터 유도된 용어의 가중치는 문장집합의 내부 특징을 잘 나타나기 때문에 문서요약의 질을 향상할 수 있다. 둘째, 가중치가 부여된 의미특징과 확장된 질의를 이용하여서 사용자의 요구사항과 제안방법의 요약결과 사이의 의미적 차이를 감소시킨다. 셋째, 하둡 기반의 3대 PC로 구성된 클러스터는 인터넷이나 소셜 네트워크 서비스의 대량의 자료를 분산저장 할 수 있어서 대량의 자료에 대한 저장문제를 해결 할 수 있다. 마지막으로, 하이브와 연구된 문서요약 알고리즘을 이용하여 분산저장된 자료를 병렬처리 하여 문서를 요약함으로서 요약의 처리속도 문제를 해결할 수 있다.

참고문헌

[1] I. Mani, M. T. Maybury, "dvances in Automatic Text," The MIT Press, 1999.

[2] A., Diaz, P., Gservas, "User-model based personalized summarization", Information Processing and Management, 43, pp.1715-1734, 2007.

[3] M., Sanderson, "Accurate user directed summarization from existing tools", In proceeding of the international conference on information and knowledge management, pp.45-51, 1998.

[4] A., Tombros, M., Sanderson, "Advantages of Query Biased summaries in Information Retrieval", In proceeding of ACM SIGIR, pp.2-10, 1998.

[5] R., Varadarajan, V., Hristidis, "A System for Query Specific Document Summarization", In proceeding of the CIKM, pp.622-631, 2006.

[6] Han, K. S., Bea, D. H., Rim, H. C., "Automatic Text Summarization Based on Relevance Feedback with Query Splitting", In proceedings of the 5th International Workshop on Information Retrieval with Asian Language, pp.201-202, 2000.

[7] 김철원, 박선, "의미특징과 워드넷 기반의 의사연관 피드백을 사용한 질의 기반의 문서요약", 한국해양정보통신학회논문지, 제15권 제7호, 2010.

[8] S. Park, D. U. An, "Automatic Query-based Personalized Summarization that uses Pseudo Relevance Feedback with NMF", In proceeding of ACM ICUIMC2010, 2010.

[9] S. Park, "User-focused Automatic Document Summarization using Non-negative Matrix Factorization and Pseudo Relevance Feedback", In proceeding of ICCEA2009, 2009.

[10] 박선, "의미 특징 행렬과 의미 가변행렬을 이용한 질의 기반의 문서 요약", 한국항행학회 논문지, 제12권, 제4호, 2008.

[11] 박선, 이주홍, "비음수 행렬 분해와 K-means를 이용한 주제기반의 다중문서요약", 한국정보과학회 논문지, 제35권, 제4호, 2008.

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