ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2013 December: 19(4): 39~53 http://dx.doi.org/10.13088/jiis.2013.19.4.039

소셜 뉴스를 위한 시간 종속적인 메타데이터 기반의 컨텍스트 공유 프레임워크

가명현
인하대학교 컴퓨터정보공학과
(gagaman7777@eslab.inha.ac.kr)
오경진
인하대학교 컴퓨터정보공학과
(okjkillo@eslab.inha.ac.kr)
홍명덕
인하대학교 컴퓨터정보공학과
(hmdgo@eslab.inha.ac.kr)
조근식
인하대학교 컴퓨터정보공학과
(gsjo@inha.ac.kr)

인터넷의 발달과 SNS의 등장으로 정보흐름의 방식이 크게 바뀌었다. 이러한 변화에 따라 소셜 미디어가 급부상하고 있으며 소셜 미디어와 비디오 콘텐츠가 융합된 소셜 TV, 소셜 뉴스의 중요성이 강조되고 있다. 이러한 환경 속에서 사용자들은 단순히 콘텐츠를 탐색만 하는 것이 아니라 같은 콘텐츠를 이용하고 있는 친구들이나 지인들과 콘텐츠에 대한 정보나 경험들을 공유하고 더 나아가 새로운 콘텐츠를 만들어내기도 한다. 하지만 기존의 소셜 뉴스에서는 이러한 사용자들의 특성을 반영해 주지 못하고 있다. 특히 이용자들의 참여성만을 고려하고 있어서 서비스간의 차별화가 어렵고 뉴스 콘텐츠에 대한 정보나 경험 공유 시 컨텍스트 공유가 어렵다는 문제가 있다. 이를 해결하기 위해 본 논문에서는 뉴스를 내용별로 분할하고 분할된 뉴스에서 추출된 시간 종속적인 메타데이터를 제공하는 프레임워크를 제안한다. 제안하는 프레임워크에서는 스토리 분할 방법을 이용하여 뉴스 대본을 내용별로 분할한다. 또한 뉴스 전체내용을 대표하는 태그, 분할된 뉴스를 나타내는 서브 태그, 분할된 뉴스가 비디오에서 시작하는 위치 즉, 시간 종속적인 메타데이터를 제공한다. 소셜 뉴스 이용자들에게 시간 종속적인 메타데이터를 제공한다면 이용자들은 전체의 뉴스 내용 중에 자신이 원하는 부분만을 탐색 할 수 있으며 이 부분에 대한 견해를 남길 수 있다. 그리고 뉴스의 전달이나 의견 공유 시 메타데이터를 함께 전달함으로써 전달하고자 하는 내용에 바로 접근이 가능하며 프레임워크의 성능은 추출된 서브 태그가 뉴스의 실제 내용을 얼마나 잘 나타내 주느냐에 따라 결정된다. 그리고 서브 태그는 스토리 분할의 정확성과 서브 태그를 추출하는 방법에 따라 다르게 추출된다. 이 점을 고려하여 의미적 유사도 기반의 스토리 분할 방법을 프레임워크에 적용하였고 벤치마크 알고리즘과 성능 비교 실험을 수행하였으며 분할된 뉴스에서 추출된 서브 태그들과 실제 뉴스의 내용을 비교하여 서브 태그들의 정확도를 분석하였다. 결과적으로 의미적 유사도를 고려한 스토리 분할 방법이 더 우수한 성능을 보였으며 추출된 서브 태그들도 컨텍스트와 관련된 단어들이 추출 되었다.

논문접수일 : 2013년 10월 25일 게재확정일 : 2013년 10월 31일
투고유형 : 국문급행 교신저자 : 조근식

1. 서론

인터넷의 발달과 SNS의 등장으로 정보흐름의 방식이 크게 바뀌었다. 단방향으로만 흐르던 정보는 양방향으로 전달되며 정보 이용자들은 단순히 정보를 수집하거나 전달 받기만 하는 것이 아니라

* 이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2011-0015484).

새로운 정보를 재생산하고 이를 다시 전파한다. 이러한 변화에 따라 소셜 미디어(Social Media)와 TV가 융합된 소셜 TV와 소셜 미디어와 뉴스가 결합된 소셜 뉴스가 급부상 하고 있다. 소셜 TV는 소셜 미디어 특유의 미디어적 속성을 고려하여, TV시청자들이 TV 수상기로 영상 콘텐츠를 시청하며 소셜 미디어를 근간으로 타인과의 소통을 통한 감정과 의견을 교환하는 형태(Yoon, 2012)를 말하며 소셜 TV의 시청 형태가 갖는 가장 보편적인 특징은 TV 프로그램 시청 중에 원격 대화 혹은 채팅을 바탕으로 동일한 프로그램을 시청하고 있는 친구들이나 지인들과 시청경험을 공유하는 것이다(Cesar and Geerts, 2011). 2012년 방송매체 이용 행태조사 보고서(Jung et al., 2012)에 따르면 국내의 TV시청자 중 17.6%가 시청중인 프로그램과 관련된 정보를 인터넷, 문자메시지, 메신저, SNS 등의 매체를 통해 검색하거나 지인들에게 전달한다고 조사되었고 이는 TV시청형태의 특징을 나타내며 소셜 TV시청형태의 특징을 설명해준다. 소셜 뉴스(Social News)는 참여형 소셜 미디어로서 뉴스의 중요도, 신뢰도가 뉴스 이용자들의 평판에 따라서 결정된다. 인터넷 미디어의 영향력 증가로 뉴스 이용자들의 관심사가 곧 사회 전반의 이슈를 나타내기 때문에 좋은 혹은 추천이 많은 뉴스들이 우선적으로 이용자에게 제공된다. 뉴스 유통방식의 이러한 변화는 뉴스 콘텐츠뿐만 아니라 뉴스의 생산관행과 뉴스 이용패턴을 변화시키고 뉴스개념에까지 영향을 미친다(Park, 2012). 많은 뉴스 서비스들이 이러한 소셜 미디어특성을 반영한 서비스를 제공하고 있다. 대표적인 상업 서비스로는 미국의 Digg[1], Reddit[2], Flipboard[3], 일본의 Gunosy[4], 그리고 국내의 위키트리[5]와 Daum view[6]가 있다. 하지만 이런 소셜 뉴스 서비스들은 뉴스 이용자들의 시청보조 수단으로만 이용되고 있다. 또한 SNS의 데이터를 분석하여 개인을 위한 뉴스를 제공하는 서비스만을 제공하고 있어 소셜 뉴스 서비스간의 차별화가 어렵다. 또 다른 문제는 소셜 TV의 시청형태에서 나타난 것처럼 뉴스 이용자는 자신의 의견을 다른 이용자들과 공유하려 하는데 이때 이용자들 간의 컨텍스트 공유에 한계가 있다. 특히 뉴스 전체를 대표하는 카테고리와 태그, 평판 등의 정보만을 제공하기 때문에 이용자가 관심이 있는 뉴스의 부분적인 내용에 접근하기 어렵고 뉴스 전체를 다 살펴보고 난 후에야 뉴스를 전달하는 사람이 말하고자 하는 바를 이해할 수 있다. 특정 경우에는 이용자들이 서로 다른 부분을 이해할 수도 있다. 정확한 컨텍스트 공유를 위해서는 기존의 소셜 뉴스 서비스에서 제공되던 정보 이외에 부가적인 정보가 필요하다.

이를 위해 본 논문에서는 뉴스 비디오의 텍스트 즉, 대본을 분석하여 뉴스의 컨텍스트별로 분할하고 분할된 각각의 뉴스에서 추출한 시간 종속적인 메타데이터를 뉴스 이용자들에게 제공하는 프레임워크를 제안한다. 뉴스는 시간 종속적인 메타데이터를 통해 더 세부적으로 분류가 되고 뉴스 이용자들은 이 부가적인 정보를 통해 뉴스를 원활하게 전달 할 수 있을 것이며 보다 정확한 컨텍스트 공유가 가능할 것이다. 그리고 이 프레임워크가 소셜 TV에도 적용이 된다면 소셜 TV시청 형태의 특징도 잘 반영 할 수 있다. 제안하는 프레임워크에서는 뉴스를 컨텍스트별로 분할하기 위해 의미

1) Digg : www.digg.com.
2) Reddit : www.reddit.com.
3) Flipboard : www.flipboard.com.
4) Gunosy(グノシー) : gunosy.com.
5) 위키트리 : www.wikitree.co.kr.
6) Daum view : v.daum.net.

적 유사도 기반의 스토리 분할 방법을 적용한다. 이 방법을 적용함으로써 기존의 어휘적 응집성 기반의 스토리 분할 방법과 달리 단어들 간의 동의어, 유의어 개념을 반영하여 스토리 분할을 할 수 있다. 그리고 프레임워크의 성능을 평가하기 위해 의미적 유사도 기반의 스토리 분할 방법과 어휘적 응집성 기반의 스토리 분할 방법 성능비교 실험을 하였고 분할된 뉴스에서 추출된 키워드들이 실제로 컨텍스트 공유에 도움이 되는지를 분석하였다.

본 논문의 구성은 다음과 같다. 제 2장에서는 제안하는 프레임워크와 관련된 배경지식과 관련 연구를 설명하고 제 3장에서는 제안하는 프레임워크의 구조와 각 구조별 기능을 설명한다. 제 4장에서는 의미적 유사도 기반의 스토리 분할 방법과 어휘적 응집성 기반의 스토리 분할 방법의 성능 비교 결과를 분석하고 컨텍스트별로 분할된 뉴스 대본에서 추출된 키워드가 실제 컨텍스트에 적합한가를 분석한다. 마지막으로 제 5장에서 결론을 맺고 향후 연구의 방향을 제시한다.

2. 배경지식 및 관련연구

2.1 문서의 단어 처리

문서에 포함된 문장들 간의 유사도를 분석하기 위해서는 문장에 등장하는 단어들의 관계파악이 필요하고 이런 단어 처리를 위해 GATE(Cunningham, 2002)와 워드넷(Miller et al., 1990)이 많이 사용된다.

GATE는 1995년 Sheffield 대학에서 개발이 시작된 자연어 처리를 위한 자바 기반의 툴이다. 주로 문서, 문장에 등장하는 단어를 추출할 때 사용되며 영어 이외에도 스페인어, 중국어, 아랍어, 프랑스어, 독일어, 이탈리아어, 러시아어 등의 자연어 처리가 가능하다. GATE는 텍스트에서의 단어 추출, 지명 추출, 문장 분할, 품사 분석 등의 기능을 제공해 준다.

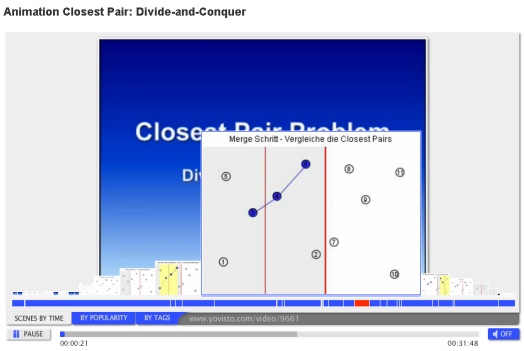
워드넷은 영어의 어휘들을 의미적 개념을 바탕으로 네트워크를 구축한 대용량 지식베이스이며 1985년부터 Princeton 대학의 G. Miller, Ch. Fellbaum 등 심리학자, 언어학자, 전산학자 등을 중심으로 구축이 시작되었다. 워드넷은 영어의 명사, 동사, 형용사와 부사를 'synset'이라는 동의어 집합으로 분류하여 간략하고 일반적인 정의를 제공한다. 또한 이 집합들이 의미적 상관관계로 연결되어 있다. WordNet::Similarity(Pedersen et al., 2004)를 활용하면 어휘들의 집합 관계에 따라 단어 간의 의미적 유사도를 구할 수 있다. 일반적으로 워드넷에 구축된 네트워크에서 유사도를 구하고자 하는 두 단어사이의 거리를 측정하고 이를 분석하여 단어간의 유사도를 구한다. 단 명사와 명사, 동사와 동사간의 유사도만 구할 수 있고 워드넷 데이터베이스에 등록되어 있지 않은 단어들의 유사도는 구할 수 없다. 많은 연구들(Budanitsky and Hirst, 2006; Pedersen et al., 2010; Oh et al., 2012)이 단어들 간의 의미적 유사도를 구하기 위해서 Word Net::Similarity를 이용하고 있다.

2.2 시간 종속적인 메타데이터

학문적 비디오 검색 엔진 서비스인 Yovisto(Sack and Waitelonis, 2010)는 비디오의 시간 종속적인 메타데이터를 제공한다. Yovisto에서는 비디오에서 발표 자료의 슬라이드가 바뀌는 이벤트가 발생하는 시점을 반자동으로 검출하여 이에 따라 학문적 비디오를 분할한다(Sack and Waitelonis, 2006). 이렇게 분할된 비디오가 시작하는 위치와 관련된 콘텐츠 즉, 시간 종속적인 메타데이터를 통해 서비

스 이용자들은 자신이 관심 있는 내용의 비디오에 바로 접근 할 수 있다. 그리고 각 내용별로 협업적 어노테이션이 가능하여 새로운 콘텐츠를 생성할 수 있다.

<Figure 1>은 Yovisto에서 비디오를 재생했을 때의 화면이다. 비디오는 강의에 사용되는 슬라이드에 따라 분할되어 있으며 분할된 비디오에 따라 나뉜 타임바(Time-bar)를 통해 서비스 이용자는 각 슬라이드가 시작하는 비디오의 위치로 바로 접근이 가능하다.



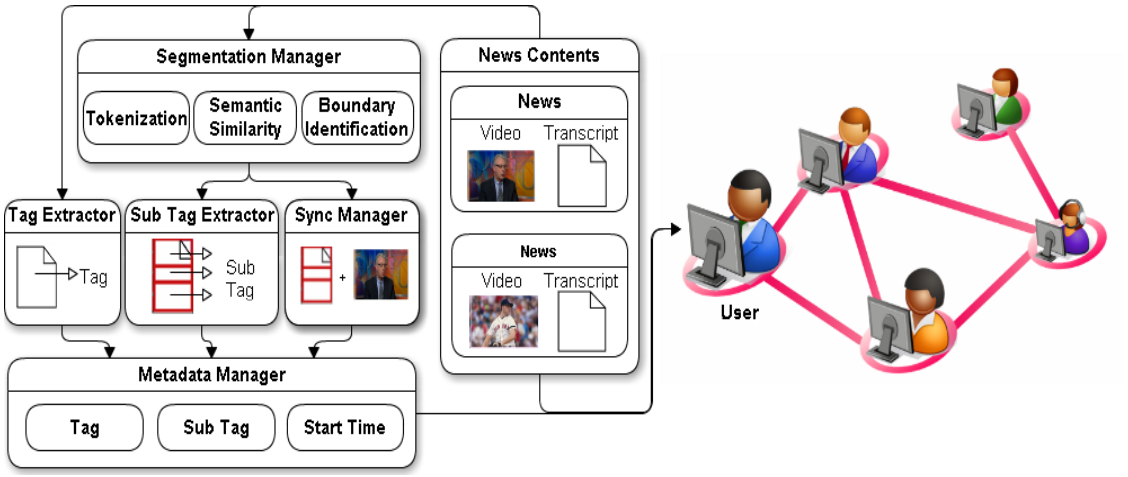
<Figure 1> Time-dependent Metadata in Yovisto

Yovisto처럼 비디오를 컨텍스트별로 분할하고 이를 뉴스이용자에게 제공한다면 뉴스이용자들도 자신이 관심 있는 뉴스를 선택하여 이용할 수 있고 사용자들 간의 정보전달이나 의견 공유시 원활한 컨텍스트 공유가 가능할 것이다. 하지만 Yovisto에서는 비디오에 등장하는 슬라이드에 따라 비디오를 분할하기 때문에 이를 바로 뉴스 서비스에 적용할 수 없고 뉴스 서비스만을 위한 별도의 시간적 메타데이터 제공방법이 요구된다.

2.3 텍스트 기반의 스토리 분할

스토리 분할(Story Segmentaion)은 텍스트나, 비디오, 오디오 등의 콘텐츠를 비슷한 의미끼리 묶어 각각의 이야기단위로 분할하는 것을 말한다. 텍스트나 비디오 같은 콘텐츠를 분할하기 위해 텍스트를 분석하거나 비디오에 등장하는 장면을 분석한다. 또한 오디오 콘텐츠를 문자음성 자동변환 기술(Text to Speech)을 사용하여 텍스트로 바꾸고 이를 분석하기도 한다. 스토리 분할의 기본 아이디어는 같은 내용을 담고 있는 부분에서 동일한 단어들이 반복되거나 유사한 이미지 패턴이 반복 될 것이고 이 정보를 바탕으로 스토리 별로 분할 한다는 것이다. 이런 스토리 분할은 주제 식별이나 내용 요약, 정보 추출, 내용별 인덱싱과 검색 등을 위해 활용되며(Lee and Chen, 2005) 정확하게 스토리를 내용별로 분할하기 위한 많은 연구들이 진행 되고 있다(Stokes et al., 2004; Misra et al., 2010; Xie et al., 2012).

대표적인 텍스트 기반의 스토리 분할 방법에는 TextTiling(Hearst, 1997)이 있다. TextTiling은 텍스트의 어휘적 응집성(Lexical Cohesion) 즉, 동일한 내용을 가지는 문장, 문단들에서 동일한 단어들이 반복적으로 등장할 것이라는 점을 고려한 스토리 분할 방법이다. 이를 위해 텍스트에서 각 단어들이 등장하는 빈도를 측정하고 공통된 단어들이 등장하는 문장, 문단들을 묶는다. TextTiling은 다음과 같은 세 가지 단계로 구성된다. 토큰화 단계에서는 텍스트의 각 문장들에 포함된 단어를 추출하고 어휘 유사도 측정 단계에서는 두 문장에 각각 등장하는 단어들의 빈도수를 측정하여 문장 간의 코사인 유사도를 계산한다. 마지막으로 스토리 변환 시점 식별 단계에서는 각 문장에서 앞뒤 문장과의 유사도 차이 정도를 의미하는 Depth Score를 계산하여 Depth Score가 높은 문장을 스토리 변환 시점으로 식별하고 이를 기준으로 전체 텍스트가 분할된다.



<Figure 2> The Structure of the Framework

3. 의미적 유사도 기반의 스토리 분할 프레임워크

<Figure 2>는 제안하는 스토리 분할 프레임워크의 구조도이고 크게 스토리 분할, 태그 추출, 서브 태그 추출, 동기화 그리고 메타데이터 관리의 5가지 모듈로 구분된다. 스토리 분할 모듈에선 토큰화, 의미적 유사도 계산, 스토리 변환시점 식별을 통해 뉴스의 대본을 스토리에 따라 분할한다. 태그 추출모듈에서는 뉴스를 대표할 수 있는 태그를 추출하고 서브 태그 추출 모듈에서는 분할된 스토리별로 컨텍스트 공유에 도움이 될 수 있는 키워드들을 추출한다. 또한 동기화 모듈에서는 분할된 대본과 비디오를 동기화 시켜 스토리 별로 비디오를 분할한다. 여기서 비디오 분할은 물리적으로 비디오를 분할하여 여러 개의 비디오 파일로 나누는 것을 의미하는 것이 아니라 각 내용들이 비디오에서 시작되는 위치를 찾아 이 위치부터 비디오를 재생하는 것을 의미한다. 메타데이터 관리에서는 뉴스 전체에서 추출된 태그들과 서브 태그 추출 모듈에서 추출된 서브 태그 그리고 동기화 모듈에서 추출된 뉴스의 각 부분이 시작되는 비디오의 위치들을 관리해주고 사용자에게 제공해준다.

3.1 의미적 유사도 기반의 스토리 분할

스토리 분할(Segmentation Manager) 모듈에서는 단어들의 동의어, 유의어 개념을 반영하기 위해 단어들의 어휘적 응집성을 고려한 유사도 측정방법을 사용하는 것이 아니라 단어들의 의미적 유사도를 고려하는 측정 방법을 사용 한다. 여기서 의미적 유사도는 두 단어가 얼마나 유사한 지를 나타내며 이는 워드넷에 구축된 네트워크에서 단어들이 떨어져 있는 정도를 분석하여 측정한다.

이를 위해 먼저 GATE를 이용하여 단어들을 추출한다. 그리고 추출된 단어들 간의 유사도를 워드넷 기반의 Wu and Palmer Similarity(Wu and Palmer, 1994) 계산 방법으로 측정한다. 그리고 측정된 단어들의 유사도를 식 (1)에 반영하여 문장 간의 의미적 유사도를 구한다(Mihalcea et al., 2006; Malik, 2007).

$$Score(q_i, q_j) = \frac{\sum_{s \in q_i} sim_m(s, q_j) + \sum_{s \in q_j} sim_m(s, q_i)}{|q_i| + |q_j|} \quad (1)$$

식 (1)에서 Score는 두 문장(q_i, q_j)간의 의미적 유사도를 나타내며 s는 각 문장에 포함된 단어를 나타낸다. 문장 q_i에 속한 단어 s와 문장 q_j에 속한 모든 단어들 간의 의미적 유사도를 측정하고 이 측정 값 중에서 제일 큰 값을 Score 계산에 반영한다. 이를 과정을 문장 q_i, q_j에 속한 모든 단어들에 대해 반복하고 두 문장에 속한 단어들의 개수로 나눈다. Score는 0부터 1까지의 값을 가지며 Score가 1이면 두 문장이 완전히 같은 의미를 지니고 있음을 뜻한다. 제안하는 프레임워크에서는 뉴스 대본에서 명사들만 추출하여 단어의 의미적 유사도를 계산한다. 또한 사람의 이름, 어느 지방의 지명과 같은 고유명사들은 워드넷 사전에 등록되어 있지 않아 고유명사간의 의미적 유사도를 측정한다고 할지라도 그 값은 0이 된다. 하지만 사람의 이름이나 지명과 같은 고유명사들은 뉴스의 컨텍스트 뉴스에서 핵심적인 단어들이기 때문에 워드넷 사전에 없는 단어의 경우에 두 문장에서 동일하게 등장한다면 두 단어 간의 의미적 유사도를 1로 설정한다.

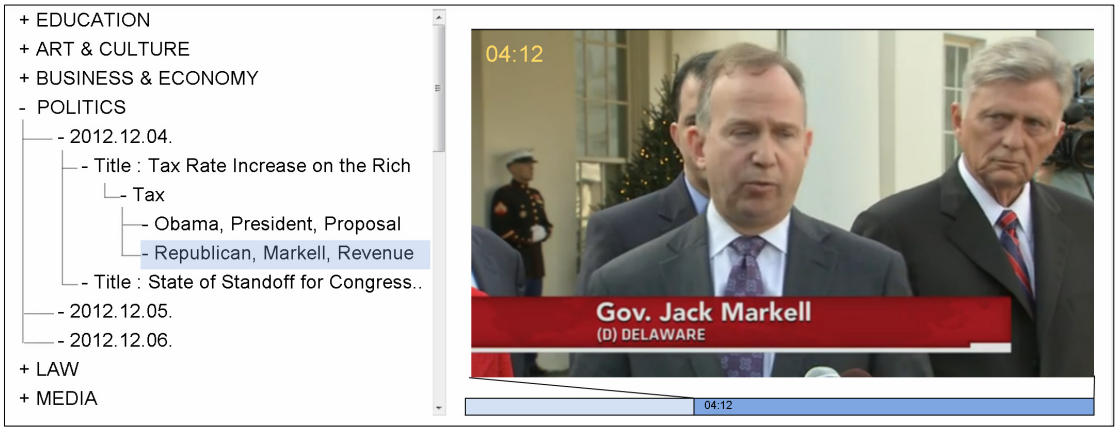
뉴스 대본에서 두 문장이 너무 멀리 떨어져 있다면 일반적으로 두 문장은 다른 내용을 뜻하는 문장이다(Xie et al., 2012). 이를 고려하기 위해 식 (2)처럼 문장연결강도를 나타내는 α값을 설정한다. 식 (2)에서 i와 j는 문장의 번호이고 α값을 1보다 작은 값으로 설정해 두면 문장이 떨어져 있는 정도에 따라 $Score(q_i, q_j)^*$의 값이 낮아지게 된다.

$$Score(q_i, q_j)^* = Score(q_i, q_j) \cdot \alpha^{|i-j|} \quad (2)$$

스토리 분할할 때 어떻게 문장 간의 유사도를 측정할 것인가도 중요한 문제이지만 문장 간의 유사도를 측정한 결과를 통해 어떻게 스토리가 변환되는 시점을 찾을 것인가도 큰 문제이다. TextTiling에서는 스토리 변환 시점 식별을 위해 각 문장에서 Depth Score를 계산하고 Depth Score가 일정 값 이상인 경우에 해당 문장을 스토리 변환 시점이라 식별한다. 그리고 이 시점에 따라 텍스트를 분할한다. 하지만 Depth Score 방식은 해당 문장에서 바로 앞뒤의 문장과의 유사도만 비교한 값으로 계산되기 때문에 같은 내용을 말하고 있는 텍스트 중간에 비유적인 표현이 사용되거나 동의어, 유의어 등이 사용된 문장이 등장한다면 정확한 스토리 분할이 어려울 수 있다. 이러한 문제를 해결하기 위해 제안하는 프레임워크에서는 스토리 변환 시점 식별 시 앞뒤 문장만 고려하는 것이 아니라 문장 전체를 고려하는 방법을 사용한다. 특정 문장과의 유사도가 일정 하한 값 이하보다 낮아지기 시작하는 문장을 찾고 이 문장을 스토리 변환 시점 후보군으로 선별한다. 그리고 뉴스의 대본에 존재하는 모든 문장에 대해 스토리 변환 시점 후보군으로 선별하고 특정횟수 이상 선별된 문장을 스토리 변활 시점이라 식별한다.

3.2 태그, 서브 태그 추출 및 동기화

제안하는 프레임워크에서는 뉴스 이용자 간의 원활한 컨텍스트 공유를 위해 뉴스 전체 내용을 대표하는 태그, 분할된 뉴스를 나타내는 서브 태그, 비디오에서 각 내용들이 시작되는 위치 즉, 시간 종속적인 메타데이터를 제공한다. 먼저 해당 뉴스와 전체 뉴스에서 단어의 TF-IDF를 측정하여 태그를 추출하고 분할된 뉴스와 해당 뉴스에서 단어의 TF-IDF를 측정하여 서브 태그를 추출한다. 그리고 분할된 각 부분이 비디오에서 시작되는 위치



<Figure 3> Example of Time-Dependent Metadata in Framework

를 구하려면 분할된 대본과 비디오를 동기화 하여야 한다. 이를 위해 Youtube API[7] 중에서 Auto synchronization를 사용한다. 이는 유튜브에 업로드 된 동영상에 자동으로 자막을 동기화하기 위한 기술이다. 동영상의 오디오 성분을 분석하여 자동으로 대본을 만들 수 있으며 이미 동영상의 자막(또는 대본)을 텍스트로 가지고 있는 경우에는 동영상의 오디오 성분과 텍스를 분석하여 동기화 된 자막파일을 제공해 주고 이 자막파일은 특정 텍스트가 등장하는 시간정보를 포함하고 있다. 뉴스 동영상과 뉴스 대본을 Youtube API를 통해 동기화 시키면 텍스트의 시간정보가 포함된 자막파일을 얻을 수 있고 이 파일의 시간정보와 분할된 뉴스 대본을 비교하여 각 분할된 뉴스가 비디오에서 시작하는 위치를 구한다.

이렇게 추출한 전체내용을 대표하는 태그, 뉴스의 각 내용을 대표하는 서브 태그 그리고 각 내용별 시작위치의 정보를 통해 뉴스 이용자들은 뉴스의 각 부분 내용을 파악하고 원하는 뉴스를 탐색한다. 또한 탐색한 정보를 다른 뉴스이용자에게 전달한다. 뉴스 전달시 이용자들은 같은 컨택스트를 이해하고 공유하여야 한다. 이를 위해 이용자들은 시간 종속적인 메타데이터를 자신이 전달하고자 하는 뉴스와 함께 전달하고 이를 전달받은 이용자는 태그와 서브태그를 통해 뉴스 전체를 살펴보지 않고도 전달받은 내용을 이해하며 내용별 시작위치 링크를 통해 전달받은 내용이 나오는 부분에 바로 접근한다. 이때 원활한 컨텍스트 공유가 가능하며 각 내용별로 협업적 어노테이션도 가능하다. 예를 들어 <Figure 3>에서 처럼 미국의 부자증세에 관한 뉴스가 있고 이 뉴스에 오바마 대통령, 민주당 의원, 공화당 의원들의 의견들이 담겨져 있을 때 시간 종속적인 메타데이터가 뉴스 이용자들에게 제공되면 뉴스 이용자는 자신이 원하는 정당의 의견만을 탐색할 수 있으며 대통령과 의원들의 의견에 자신의 견해를 해당 뉴스에 덧붙이거나 다른 이용자들에게 전달할 수 있다. 또한 뉴스를 전달받은 이용자들은 뉴스 전체에서 추출된 태그인 'Tax'와 각 내용에서 추출된 'Obama', 'President', 'Repeblican', 'Markell'와 같은 서브 태그를 참조하여 뉴스 전달자가 전달하고자 하는 바를 보다 원활하게 파악할 수 있다.

7) Youtube API : https://developers.google.com.

4. 실험 및 분석

제안하는 프레임워크는 뉴스 이용자 간의 원활한 컨텍스트 공유를 위해 뉴스 비디오를 분할하고 이 분할된 부분에서 시간 종속적인 메타데이터를 추출하여 사용자에게 제공하는데 목적을 두고 있다. 이 때문에 프레임워크의 성능은 스토리 분할의 정확도와 추출되는 태그와 서브 태그들의 결과에 따라 평가된다. 이 점을 고려하여 다음과 같이 두 가지 결과를 평가하고 분석하였다. 먼저 프레임워크에서 이용한 의미적 유사도 기반의 스토리 분할 방법의 성능을 알아보기 위해 어휘적 응집성 기반의 스토리 분할 방법인 TextTiling과 비교 실험을 진행하였고 그 결과를 분석하였다. 또한 뉴스에서 추출된 서브 태그들이 실제로 이용자들 간의 컨텍스트 공유에 도움이 되는지를 평가하기 위해 뉴스의 실제 내용과 추출된 서브 태그를 분석하였다.

제안하는 프레임워크의 성능 실험 위해 PBS[8]에서 방영된 뉴스 중 PBS NewsHour[9]를 이용하였다. 미국의 공영 방송사인 PBS는 미국의 비영리 민간법인 방송국이며, 1970년에 교육목적의 방송국으로 개국하였고 1973년에 종합방송국으로 전환 되었다. PBS NewsHour에서 총 93편(2012년 12월 3일부터 2012년 12월 18일까지)의 뉴스를 수집하였다. PBS NewsHour는 한 시간 분량의 일간 뉴스이며 다양한 카테고리의 뉴스를 제공한다. 또한 PBS NewsHour에서는 뉴스의 비디오와 뉴스의 대본을 함께 제공하기 때문에 실험에 사용하기 적합하다. 실험을 위해 3명의 평가자가 직접 스토리별로 분할하였고 추출된 서브 태그와 스토리별 내용을 비교하였다.

8) PBS(Public Broadcasting Service) : www.pbs.org.
9) PBS NewsHour : www.pbs.org/newshour.

4.1 스토리 분할 방법 성능 비교 평가

제안하는 프레임워크에서 사용한 의미적 유사도 기반의 스토리 분할 방법은 TextTiling 방법과 크게 2가지 점이 다르다. 먼저 문장 간의 유사도를 구하는 방법이 다르다. 두 가지 스토리 분할 방법 모두 문장 속에 포함된 단어들을 가지고 문장을 구하지만 TextTiling에서는 단어들의 빈도수 즉, 어휘적 응집성만을 계산하여 문장 간의 유사도를 구한다. 이와 달리 제안하는 방법에서는 단어들의 의미적 유사도를 계산하여 문장 간의 유사도를 구한다. 다른 차이점은 스토리 변환 시점을 식별하는 방법이다. TextTiling에서는 각 문장에서 앞뒤 문장과의 유사도만을 반영하여 스토리 변환 시점을 식별하는 반면 제안하는 스토리 분할 시점 식별 방법에서는 모든 문장 간의 유사도를 반영하기 위해 특정 문장에서 유사도가 낮아지기 시작하는 문장을 찾고 이를 변환 시점 후보군을 선별하고 모든 문장에 대해서 후보군을 선별한 뒤 선별된 횟수가 높은 문장을 스토리 변환 시점이라 식별한다. 제안하는 스토리 분할방법의 성능과 스토리 변환 시점 식별 방법의 성능을 각각 알아보기 위해 다음의 세 가지 방법을 통해 성능 비교 실험을 진행하였다.

① TextTiling
② 문장의 의미적 유사도+Depth Score
③ 문장의 의미적 유사도+후보군 선별법

그리고 문장연결강도를 나타내는 α값은 실험을 통해 0.98로 설정하였다. α값이 0.98보다 적은 경우에는 문장의 유사도들이 너무 낮게 나와 스토리 분할이 어려웠다.

스토리 분할 시 가장 큰 문제는 스토리가 몇 개로 분할될지 모른다는 것이다. 스토리가 몇 개로 분할되는지 미리 알고 있다면 스토리 분할 정확도

는 비약적으로 증가할 것이다. 하지만 뉴스가 몇 개의 스토리를 가지는지 알 수 없고 스토리의 개수는 뉴스마다 다 다르다. 이를 위해 동적인 스토리 변환 시점 식별 방법이 필요하다. 후보군 선별을 통한 분할 방법에서는 문장 간의 유사도가 0.4 이하로 낮아지기 시작하는 두 문장을 스토리 후보군을 선별하고 평균보다 더 많이 후보군으로 선별된 문장을 스토리 식별 시점으로 식별한다. 이 때 너무 많은 스토리 변환 시점이 식별되어 뉴스가 너무 잘게 분할되는 경우를 피하기 위해서 후보군으로 선별된 횟수가 3번 이하인 경우는 후보군 평균 계산에서 제외한다.

<Table 1> F1-Score

	Avg.	Std. Dev.
TextTiling	0.7070	0.1438
Semantic+Depth Score	0.7829	0.1330
Semantic+Candidates	0.8403	0.0776

세 가지 자동 스토리 분할 방법으로 뉴스를 분할하여 같은 내용을 담고 있다고 판단되는 문장들을 하나의 집합으로 만들었고 모든 뉴스에 존재하는 모든 문장 집합에서의 F1-Score를 측정하여 평균과 표준편차를 구하였고 그 결과는 <Table 1>과 같다. 문장 간의 유사도를 구할 때 어휘적 응집성을 고려하는 방법보다 의미적 유사도를 고려하는 방법이 10.73% 더 정확하였다. 또한 스토리를 분할 할 때 Depth Score 방식보다 후보군 선별법이 7.34% 정확했다. 문장의 의미적 유사도와 후보군 선별 두 가지를 다 적용하여 스토리 분할을 한 경우에는 TextTiling보다 18.85% 정확했다. 또한 표준편차도 기존의 방법보다 더 낮게 나타났고 이는 모든 뉴스에 대해 균일한 정도의 F1-Score가 측정되었음을 의미한다. 하지만 세 가지 방법에서 공통적으로 문장이 짧아 유사도 계산에 활용할 단어가 적은 경우 스토리 분할이 정확하게 되지 않는 문제가 발생하였다. 이 문제는 자연어 처리에서 많이 발생하는 문제이다. 어떤 특정 문장이 스토리 분할 시점이라고 식별된 경우를 살펴보면 스토리 분할 시점이 아니지만 문장이 짧아 잘못 식별된 경우가 발생하였다.

4.2 추출된 서브 태그 평가

분할된 뉴스에서 추출된 서브 태그들을 뉴스 이용자에게 제공하기 때문에 이 서브 태그들이 얼마나 컨텍스트 공유에 도움이 되는지 또한 프레임워크의 성능을 의미하게 된다. 이를 위해 문장의 의미적 유사도와 스토리 분할 시점의 후보군 선별을 모두 고려하여 스토리 분할을 한 뒤 분할된 뉴스에서 서브 태그를 추출하고 이 서브 태그들이 실제로 컨텍스트와 얼마나 관련이 있는지를 분석하였으며 그 결과는 <Table 2>와 같다.

<Table 2> Evaluate the Accuracy of Sub Tag

	Basic		Applying Stopwords	
	Avg.	Std. Dev.	Avg.	Std. Dev
Top 3	0.565	0.198	0.648	0.193
Top 5	0.528	0.166	0.739	0.175

<Table 2> 평가자들이 자동으로 추출된 서브 태그 들 중에서 실제 컨텍스트와 관련이 있는지를 평가하고 그 비율을 나타낸 것이다. 표에서 Basic은 불용어를 적용하지 않았을 때이고 Avg.는 서브 태그의 평균적인 정확도이며 Std. Dev.는 표준편차이다. Top3는 단어들의 TF-IDF를 계산하였을 때 상위 3개의 단어를 서브 태그로 추출한 경우를 말하고 Top5는 상위 5개의 단어를 서브 태그로 추

출한 경우를 말한다. 그리고 서브 태그분석에서 적용한 불용어는 PBS NewsHour의 앵커들과 리포터들의 이름이다. 실험에 사용한 뉴스 데이터는 해당 문장을 말하는 화자의 정보도 포함되어 있다. 즉 대본에 앵커나 리포터의 이름이 자주 등장하는데 이는 뉴스의 내용과는 상관없이 등장한다. 이 때문에 앵커의 이름과 리포터의 이름을 불용어로 지정하고 다시 서브 태그를 추출하였다. 이와 달리 뉴스에 등장하는 화자의 이름은 컨텍스트와 관련된 내용이다. 예를 들어 미국 대통령의 이름이나 미국 의원들의 이름은 분할된 각 부분의 내용과 연관되어있다. 평균적으로 추출된 서브 태그는 불용어를 적용하지 않은 경우 Top3와 Top5에서 각각 56.5%, 52.8%의 비율로 해당 뉴스와 관련이 있는 단어들 이었고 불용어를 적용한 경우에는 Top3와 Top5에서 각각 64.8%, 74.4%의 비율로 해당 뉴스와 관련된 내용들이 서브 태그로 추출되었다. 추출된 서브 태그의 정확도에 대한 표준편차는 Top3의 경우에 0.198에서 0.193으로 감소하였고 Top5의 경우에 0.166에서 0.175로 증가하였다.

불용어를 적용하여 자동으로 서브 태그를 추출하였을 때 서브 Top3 중에서는 2개 정도의 단어가 Top5 중에서는 3.7개 정도의 단어가 컨텍스트와 관련이 있었다. 그리고 컨텍스트와 관련이 없는 단어들이 추출된 원인을 찾기 위해 실험결과를 분석하였고 분석 결과는 아래와 같다.

먼저 두 단어(또는 형태소)가 하나의 뜻을 이루고 있을 때 두 단어를 나누어 빈도수를 측정하기 때문에 정확한 서브 태그를 추출할 수 없었다. 뉴스 대본에 자주 등장하는 'house'의 경우 '집'이라는 의미로 사용된 경우 보다는 'White house'가 두 단어로 분리 되고 따로 처리가 되었을 때 등장하는 경우가 더 많았고 'White house' 대신에 'house'가 서브 태그로 추출이 되었다. 정확한 서브 태그의 추출이 어려운 또 다른 이유는 서브 태그로 명사만을 추출하였기 때문이었다. 일반적으로 태그나 키워드는 콘텐츠의 내용을 대표하는 명사들이다. 스토리 분할을 위해 의미적 유사도를 계산할 때는 문장 중에서 명사만을 추출하여도 실험 결과에서처럼 어느 정도 정확한 스토리 분할을 할 수 있었지만 해당 콘텐츠의 정확한 컨텍스트를 파악하기 위해선 동사나 형용사의 분석도 필요하다. 명사를 이용하는 경우 어느 특정 인물(예 : 대통령, 의원)이나 장소와 관련된 내용에 대해서는 비교적 정확한 정보를 찾을 수 있었지만 명사들과 관련된 동사에 대한 정보가 빠져있어 해당 인물이 말하고자 하는 바를 파악하기 어려웠고 형용사 정보도 고려하지 않았기 때문에 이 인물들의 말한 내용이 긍정적인 내용인지 부정적인 내용인지 알 수가 없었다. 정확한 키워드 추출이 어려운 또 다른 이유로는 너무 일반적인 단어들이 서브태그로 추출이 된다는 점이었다. 'way', 'issue', 'kind'와 같은 단어들은 빈도수가 높아 서브 태그로 추출이 되었지만 영어 문장에서 너무 자주 사용되는 단어여서 각 내용의 컨텍스트와 관련짓기가 어려웠다. 마지막으로 같은 의미를 가지는 단어이지만 서로 다른 표현을 사용했을 때 정확한 서브 태그 추출이 어려웠다. 예로 든 뉴스에서 'Obama'와 'President'는 같은 인물을 지칭하지만 두 단어는 다른 단어로 인식되며 워드넷을 활용하여도 'Obama'라는 단어가 워드넷 사전에 등록되어있지 않기 때문에 두 단어가 같은 단어라고 처리할 수가 없었다. 이러한 점들 때문에 <Table 2>에서처럼 키워드의 정확도가 낮게 나타났다.

5. 결론 및 향후 연구

본 논문에서는 소셜 뉴스 서비스와의 차별성을

부여하고 사용자들의 TV시청 형태를 고려하여 뉴스 전달시 보다 원할한 컨텍스트 공유가 가능하게 하기위해 뉴스를 내용별로 분할하고 분할된 뉴스에서 시간 종속적인 메타데이터를 제공하는 프레임워크를 제안하였다. 그리고 이 프레임워크에서는 의미적 유사도 기반의 스토리 분할 방식을 이용하였다.

프레임워크의 성능은 스토리 분할과 추출된 서브 태그의 정확도에 따라 평가된다. 이 점을 고려하여 기존의 어휘적 응집성 기반의 스토리 분할 방법과 성능 비교 실험을 하였고 프레임워크에 적용한 의미적 유사도 기반의 스토리 분할 방식의 정확도가 18.85% 만큼 높았다. 또한 추출된 서브 태그들이 실제 내용과 관련이 있는지 분석한 결과 3개의 서브 태그를 추출하였을 때는 평균적으로 64.8% 만큼, 5개의 서브 태그를 추출하였을 때는 평균적으로 73.9% 만큼 내용과 관련이 있는 단어들이 서브 태그로 추출되었다. 그리고 스토리 분할의 경우 문장이 짧은 경우에 스토리 분할이 정확하게 되지 않는 문제가 발생하였고 자연어 처리의 어려움으로 인해 각 내용을 대표하는 서브 태그들을 정확히 추출하기 어려웠다. 하지만 시간 종속적인 메타데이터는 뉴스 콘텐츠 이용자들은 원활하게 관심이 있는 내용 탐색이 가능하여 소셜 뉴스의 특징인 참여성을 보다 향상시킬 수 있다. 그리고 태그와 서브 태그는 뉴스 관련 정보 전달시 뉴스 이용자 간의 컨텍스트 공유에도 도움을 줄 수 있다.

위에서 언급한 오차를 줄이고 보다 정확한 프레임워크 설계를 위해 스토리 분할에서 공통적으로 등장하는 문제해결에 대한 향후 연구가 요구되며 정확한 서브 태그를 추출할 수 있는 연구도 필요하다. 소셜 뉴스 뿐만 아니라 비디오 콘텐츠를 제공하는 다른 서비스에서도 시간 종속적인 메타데이터의 장점을 적용하는 방법에 대한 연구도 진행이 된다면 비디오 콘텐츠와 소셜 미디어가보다 융합되는 매개점이 될 것이다.

참고문헌

Budanitsky, A. and G. Hirst, “Evaluating wordnet-based measures of lexical semantic relatedness,” *Computational Linguistics*, Vol. 32, No.1(2006), 13~47.

Cesar, P. and D. Geerts, “Past, present, and future of social TV : A categorization,” *In Consumer Communications and Networking Conference*, (2011), 347~351.

Cunningham, H., “GATE, a general architecture for text engineering,” *Computers and the Humanities*, Vol.36, No.2(2002), 223~254.

Hearst, M. A., “TextTiling : Segmenting text into multi-paragraph subtopic passages,” *Computational Linguistics*, Vol.23, No.1(1997), 33~64.

Jung, Y.-C., N.-D. Kim and Y.-H. Kim, 2012 *Broadcast Media Usage Patterns Research*, Korea communications commission, 2012.

Lee, L. S. and B. Chen, “Spoken document understanding and organization,” *IEEE Signal Processing Magazine*, Vol.22, No.5(2005), 42~60.

Mihalcea, R., C. Corley and C. Strapparava, “Corpus-based and knowledge-based measures of text semantic similarity,” *Proceedings of the American Association for Artificial Intelligence*, Vol.6(2006), 775~780.

Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross and K. J. Miller, “Introduction to wordnet : An on-line lexical database,” *International Journal of Lexicography*, Vol.3, No.4(1990), 235~244.

Misra, H., F. Hopfgartner, A. Goyal, P. Punitha and J. M. Jose, "Tv news story segmentation based on semantic coherence and content similarity," *Advances in Multimedia Modeling*, Vol.5916(2010), 347~357.

Oh, J. H., K. Torisawa, C. Hashimoto, T. Kawada, S. De Saeger, J. I. Kazama and Y. Wang, "Why question answering using sentiment analysis and word classes," *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics*, (2012), 368~378.

Park, S.-H., "SNS News communicating," *Communication and Information Research*, Vol. 49, No.2(2012), 37~73.

Pedersen, T., S. Patwardhan and J. Michelizzi, "WordNet::Similarity : measuring the relatedness of concepts," *Demonstration Papers at HLT-NAACL 2004, Association for Computational Linguistics*, (2004), 38~41.

Pedersen, T., "Information content measures of semantic similarity perform better without sense-tagged text," *In Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics*, (2010), 329~332.

R Malik, L. V. Subramaniam, and S. Kaushik, "Automatically Selecting Answer Templates to Respond to Customer Emails," *In Proceedings of the 20th international joint conference on Artifical intelligence*, Vol.7(2007), 1659~1664.

Sack, H. and J. Waitelonis, "Automated annotations of synchronized multimedia presentations," *In Proceedings of the ESWC 2006 Workshop on Mastering the Gap : From Information Extraction to Semantic Representation, CEUR Workshop Proceedings*, (2006).

Sack, H. and J. Waitelonis, "Exploratory Semantic Video Search with yovisto," *In Semantic Computing (ICSC), IEEE Fourth International Conference on IEEE*, (2010), 446~447.

Stokes, N., J. Carthy and A. F. Smeaton, "SeLeCT : a lexical cohesion based news story segmentation system," *AI Communications*, Vol.17, No.1(2004), 3~12.

Wu, Z. and M. Palmer, "Verb semantics and lexical selection," *Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics*, (1994), 133~138.

Xie, L., L. Zheng, Z. Liu and Y. Zhang, "Laplacian eigenmaps for automatic story segmentation of broadcast news," *Audio, Speech, and Language Processing, IEEE Transactions*, Vol. 20 No.1(2012), 276~289.

Yoon, S.-H., *Revolution of Social TV*, ebizbooks, 2012.

Abstract

Context Sharing Framework Based on Time Dependent Metadata for Social News Service

Myung-Hyun Ga* · Kyeong-Jin Oh* · Myung-Duk Hong* · Geun-Sik Jo**

The emergence of the internet technology and SNS has increased the information flow and has changed the way people to communicate from one-way to two-way communication. Users not only consume and share the information, they also can create and share it among their friends across the social network service. It also changes the Social Media behavior to become one of the most important communication tools which also includes Social TV. Social TV is a form which people can watch a TV program and at the same share any information or its content with friends through Social media. Social News is getting popular and also known as a Participatory Social Media. It creates influences on user interest through Internet to represent society issues and creates news credibility based on user's reputation. However, the conventional platforms in news services only focus on the news recommendation domain. Recent development in SNS has changed this landscape to allow user to share and disseminate the news. Conventional platform does not provide any special way for news to be share. Currently, Social News Service only allows user to access the entire news. Nonetheless, they cannot access partial of the contents which related to users interest. For example user only have interested to a partial of the news and share the content, it is still hard for them to do so. In worst cases users might understand the news in different context. To solve this, Social News Service must provide a method to provide additional information. For example, Yovisto known as an academic video searching service provided time dependent metadata from the video. User can search and watch partial of video content according to time dependent metadata. They also can share content with a friend in social media. Yovisto applies a method to divide or synchronize a video based whenever the slides presentation is changed to another page. However, we are not able to employs this method on news video since the news video is not incorporating with any power point slides presentation. Segmentation method is required to separate the news video and to creating time dependent metadata. In this work, In this

* Department of Computer and Information Engineering, Inha University
** Corresponding Author: Geun-Sik Jo
School of Computer and Information Engineering, Inha University
100 inharo, Nam-gu, Incheon 402-751, Korea
Tel: +82-32-875-5863, Fax: +82-32-875-5863, E-mail: gsjo@inha.ac.kr

paper, a time dependent metadata-based framework is proposed to segment news contents and to provide time dependent metadata so that user can use context information to communicate with their friends. The transcript of the news is divided by using the proposed story segmentation method. We provide a tag to represent the entire content of the news. And provide the sub tag to indicate the segmented news which includes the starting time of the news. The time dependent metadata helps user to track the news information. It also allows them to leave a comment on each segment of the news. User also may share the news based on time metadata as segmented news or as a whole. Therefore, it helps the user to understand the shared news. To demonstrate the performance, we evaluate the story segmentation accuracy and also the tag generation. For this purpose, we measured accuracy of the story segmentation through semantic similarity and compared to the benchmark algorithm. Experimental results show that the proposed method outperforms benchmark algorithms in terms of the accuracy of story segmentation. It is important to note that sub tag accuracy is the most important as a part of the proposed framework to share the specific news context with others. To extract a more accurate sub tags, we have created stop word list that is not related to the content of the news such as name of the anchor or reporter. And we applied to framework. We have analyzed the accuracy of tags and sub tags which represent the context of news. From the analysis, it seems that proposed framework is helpful to users for sharing their opinions with context information in Social media and Social news.

Key Words : Context Sharing, Story Segmentation, Semantic Similarity, Social Media, Social News

저 자 소 개

Myung-Hyun Ga

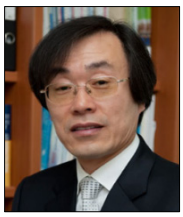
Received a B.S. degree Computer and Information Engineering from Inha University, Korea in 2012. he is currently M.S. Candidate in Information Engineering of Inha University, Korea. His research interests include Recommender System and Data Mining.

Kyeong-Jin Oh

Received a B.S. degree Computer and Information Engineering from Inha University, Korea, in 2006, and a M.S. degree in Information Engineering from Inha University, Korea in 2008. He is a Ph.D. Candidate in Information Engineering of Inha University, Korea. His research interests include Data Mining and Semantic Web.

Myung-Duk Hong

Received a B.S. degree Computer Science, Seoul Digital University, Korea, in 2008, and a M.S. degree in Information Engineering from Inha University, Korea, in 2011. He worked for Kuwoo Information Technology as a researcher 2005 to 2008. He is currently a Ph.D. Candidate in Information Engineering of Inha University, Korea. His research interests include Recommender System, Semantic Web, Ant Colony Optimization and Meta-Heuristic.



Geun-Sik Jo

Is a Professor in Computer and Information Engineering, Inha University, Korea. He received the B.S. degree in Computer Science from Inha University in 1982. He received the M.S. and the Ph.D. degrees in Computer Science from City University of New York in 1985 and 1991, respectively. He has been the General Chair and/or Technical Program Chair of more than 20 international conferences and workshops on artificial intelligence, knowledge management, and semantic applications. His research interests include knowledge-based scheduling, ontology, semantic Web, intelligent E-Commerce, constraint-directed scheduling, knowledge-based systems, decision support systems, and intelligent agents. He has authored and coauthored five books and more than 200 publications.