ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2014 June: 20(2): 39~58 http://dx.doi.org/10.13088/jiis.2014.20.2.039

# 텍스트 데이터 시각화를 위한 MVC 프레임워크

**최광선**
솔트룩스 전략사업본부
(*kschoi@saltlux.com*)

**정교성**
솔트룩스 전략사업본부
(*ksjeong@saltlux.com*)

**김수동**
숭실대학교 컴퓨터학과
(*sdkim777@gmail.com*)

빅데이터의 중요성에 대한 인식이 확산되고, 관련한 기술이 발전됨에 따라, 최근에는 빅데이터의 처리와 분석의 결과를 어떻게 시각화할 것인지가 매우 관심 받는 주제로 부각되고 있다. 이는 분석된 결과를 보다 명확하고 효과적으로 전달하는 데에 있어서 데이터의 시각화가 매우 효과적인 방법이기 때문이다. 시각화는 분석 시스템과 사용자가 소통하기 위한 하나의 그래픽 사용자 인터페이스(GUI)를 담당하는 역할을 한다. 통상적으로 이러한 GUI 부분은 데이터의 처리나 분석의 결과와 독립될 수록 시스템의 개발과 유지보수가 용이하며, MVC(Model-View-Controller)와 같은 디자인 패턴의 적용을 통해 GUI와 데이터 처리 및 관리 부분 간의 결합도를 최소화하는 것이 중요하다. 한편 빅데이터는 크게 정형 데이터와 비정형 데이터로 구분할 수 있는데 정형 데이터는 시각화가 상대적으로 용이한 반면, 비정형 데이터는 시각화를 구현하기가 복잡하고 다양하다. 그럼에도 불구하고 비정형 데이터에 대한 분석과 활용이 점점 더 확산됨에 따라, 기존의 전통적인 정형 데이터를 위한 시각화 도구들의 한계를 벗어나기 위해 각각의 시스템들의 목적에 따라 고유의 방식으로 시각화 시스템이 구축되는 현실에 직면해 있다. 더욱이나 현재 비정형 데이터 분석의 대상 중 대부분을 차지하고 있는 텍스트 데이터의 경우 언어 분석, 텍스트 마이닝, 소셜 네트워크 분석 등 적용 기술이 매우 다양하여 하나의 시스템에 적용된 시각화 기술을 다른 시스템에 적용하는 것이 용이하지 않다. 이는 현재의 텍스트 분석 결과에 대한 정보 모델이 서로 다른 시스템에 적용될 수 있도록 설계되지 못하는 경우가 많기 때문이다. 본 연구에서는 이러한 문제를 해결하기 위하여 다양한 텍스트 데이터 분석 사례와 시각화 사례들의 공통적 구성 요소들을 식별하여 표준화된 정보 모델인 텍스트 데이터 시각화 모델을 제시하고, 이를 통해 시각화의 GUI 부분과 연결할 수 있는 시스템 모델로서의 시각화 프레임워크인 TexVizu를 제안하고자 한다.

**주제어** : 텍스트 데이터, 시각화, MVC 프레임워크

논문접수일 : 2014년 3월 27일 논문수정일 : 2014년 4월 28일 게재확정일 : 2014년 5월 12일
투고유형 : 국문급행 교신저자 : 김수동

## 1. 서론

빅데이터의 중요성에 대한 인식이 확산되고, 관련한 기술이 발전됨에 따라, 최근에는 빅데이터의 처리와 분석의 결과를 어떻게 시각화할 것인지가 매우 관심 받는 주제로 부각되고 있다. 이는 분석된 결과를 보다 명확하고 효과적으로 전달하는 데에 있어서 데이터의 시각화가 매우 효과적인 방법이기 때문이다.

그런데 시각화는 분석 시스템과 사용자가

* 본 논문은 미래창조과학부 산업원천기술개발사업(10044494, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)의 지원으로 수행되었음

소통하기 위한 하나의 그래픽 사용자 인터페이스(GUI)를 담당하는 역할을 한다. 통상적으로 이러한 GUI 부분은 데이터의 처리나 분석의 결과와 독립될 수록 시스템의 개발과 유지보수가 용이하다. 그리고 이를 구현하기 위해서는 GUI와 데이터 처리 및 관리 부분 간의 결합도를 최소화하는 것이 중요하며, MVC(Model-View-Controller)라는 디자인 패턴을 이러한 문제를 해결하기 위한 중요한 디자인 기법이다 (Gamma et al., 1995).

한편 빅데이터는 크게 정형 데이터와 비정형 데이터로 구분할 수 있는데, 여전히 데이터 분석의 중심이 되는 정형 데이터의 경우는 대상 데이터나 분석하는 방법과 결과 중 많은 부분이 일정한 패턴에 따라 시각화되는 경우가 많아 MVC 디자인 패턴의 적용이 상대적으로 용이하며, 이를 적용한 다양한 시각화 도구들이 상용 수준으로 패키지화되어 제공되고 있다.

그러나 비정형 데이터에 대한 분석과 활용이 점점 더 확산됨에 따라, 기존의 전통적인 정형 데이터를 위한 시각화 도구들의 한계를 벗어나기 위해 각각의 시스템들의 목적에 따라 고유의 방식으로 시각화 시스템이 구축되는 현실에 직면해 있다. 특히나 현재 비정형 데이터 분석의 대상 중 대부분을 차지하고 있는 텍스트 데이터의 경우 언어 분석, 텍스트 마이닝, 소셜 네트워크 분석 등 적용 기술이 매우 다양하여 하나의 시스템에 적용된 시각화 기술을 다른 시스템에 적용하는 것이 용이하지 않다. 이는 현재의 텍스트 분석 결과에 대한 정보 모델이 하나의 시스템에 적용한 후 목적이 다른 또 다른 시스템에 확장 적용하는 것을 고려하여 설계되지 못하는 경우가 많기 때문이다.

본 연구에서는 이러한 문제를 해결하기 위하여 다양한 형태의 텍스트 데이터 분석 사례와 시각화 사례들의 공통적 구성 요소들을 식별하여 표준화된 형태의 정보 모델인 텍스트 데이터 시각화 모델을 제시하고, 이를 통해 시각화의 GUI 부분과 연결할 수 있는 시스템 모델로서의 시각화 프레임워크인 TexVizu를 제안하고자 한다.

이를 위해 본 논문의 2장 ‘관련연구'에서는 텍스트 데이터 시각화에 대한 관련사례들을 검토하고 다양한 텍스트 사례들 중 각각의 특징을 가지는 대표 사례를 조사하며, 3장 '텍스트 데이터 시각화 프레임워크’에서는 다양한 시각화 사례들을 구조적 관점과 상호작용 관점, 내용적 관점에서 분석하여 텍스트 데이터 시각화를 위한 공통 정보 모델을 구성하고, 이를 통해 텍스트 데이터 시각화를 위한 시스템적 공통요구사항을 도출하며, 이를 이용하여 시스템적 프레임워크인 TexVizu를 제시한다. 4장 ‘사례연구’에서는 제시된 TexVizu 프레임워크를 이용하여, 이미 연구진에 의해 구현된 두 가지 대표사례에 대한 논리적 적용을 통해 제시된 시각화 프레임워크의 적용 가능성을 검증한다. 마지막으로 5장 ‘결론’에서는 현재 연구에 대한 의미와 향후 연구 방향에 대한 소개를 통해 결론을 맺고자 한다.꼬리말

## 2. 관련연구

일반적으로 텍스트 시각화는 데이터 시각화 연구의 한 분야로, 텍스트 데이터를 시각적으로 표현함으로써 그 데이터 내부에 숨은 정보를 보다 쉽게 파악하는 것을 목적으로 한다. 텍스트 시각화는 일차적으로 텍스트 데이터가

갖는 텍스트의 형식, 구조, 내용, 패턴, 관계, 트렌드 등을 직관적으로 전달한다. 또한 시각적 가공을 통해 스토리텔링 창출을 위한 소재로 사용하거나 예술 작품을 위한 소재로 활용하는 사례를 찾아볼 수 있다(Kim and Park, 2013).

특히 텍스트 시각화의 효과 관련 분야에서 김효영, 박진완의 “텍스트 데이터의 특성에 따른 성경 시각화 사례 분석”연구는 성경 데이터 기반의 다양한 시각화의 사례 분석을 통해 텍스트가 갖는 다양한 특성에 따른 시각화 접근 방식에 대한 이론적 토대를 마련하였다. 텍스트가 가지는 내용적, 구조적 특성 및 인용정보를 텍스트의 특성으로 정의하고, 각 특성에 적합한 텍스트 시각화 사례를 분석하였다. 이를 통해 효과적인 시각화를 설계 또는 해석하기 위해서는 시각화의 재료가 되는 텍스트 데이터에 대한 체계적인 분석이 필요함을 주장하였다(Kim and Park, 2013). 또한, 이지연은 사용자에게 텍스트 데이터에 대한 시각적인 표현을 제공하는 것의 이용성 평가를 통해 텍스트 데이터의 시각화 표현에 대한 고려 사항을 발견하였다(Jones et al., 2013). 한편, 이진희는 시각화 요소에 대한 효용성 분석을 통해 데이터의 유형에 따른 사용자들의 시각화 요소를 제안하였다. 시각화 요소를 네트워크 브라우저, 그래픽 차트, 텍스트 리스트, 지형도로 분류하고 각 분류에 대한 효용성 평가를 수행하였다. 시각화 요소의 효용성은 6가지로 정의하였으며, 각 효용성에 대한 평가 기준을 제시하였다(Lee, 2005). Anton Heijs는 그의 연구에서 텍스트 데이터 분석에 있어서 텍스트 데이터 시각화의 중요성을 재강조하였다. 특히 텍스트 데이터 시각화는 데이터의 복잡한 패턴을 이해하거나 의사 결정 지원에 있어서 그 유용함에 대해 보여 주었다(Gansner et al.,2013).

텍스트 분석 방법과 시각화 방안과 관련하여 의료정보학 분야에서는 의학 논문의 텍스트를 추출, 분석, 시각화하는 다양한 연구가 진행되고 있다. Jahiruddin은 의학 텍스트로부터 의학적 개념들을 추출하고 기존 의학적 지식과 통합하여 시각화하는 BioKEVis를 개발하였다. 이 연구를 통해 연구자들은 의학 텍스트 내의 주요 의학적 개념들을 쉽게 이해하고, 텍스트 간의 의미적 관계를 파악할 수 있게 되었다(Jahiruddin et al., 2010). 또한, Jones는 텍스트 시각화가 프로젝트 관리를 효율적으로 지원할 수 있다고 주장하였다. 프로젝트 관련 이메일을 분석하여 프로젝트의 세부 항목 별 긍정, 부정을 파악하여 프로젝트 전반뿐 아니라 세부 항목 별 현황을 파악이 가능함을 제시하고 있다(Jones et al., 2013). Zhao는 소셜 텍스트 스트림으로부터 이벤트를 감지하고 시각화하는 연구를 통해 특정 주제, 관계 인물, 그리고 시간 간의 관계를 보다 효율적으로 파악하는 방법을 제시하였다. 이 방법에서는 이벤트를 인물, 토픽, 그리고 시간의 관계로 정의하고, 각 이벤트를 내용 차원, 시간 차원, 소셜 차원의 3차원으로 표현하여 소셜 텍스트 스트림에 내포된 이벤트를 시각화고 있다(Zhao and Mitra, 2007). 한편, Emden R. Gansner는 그의 연구에서 Dynamic Maps를 활용하여 텍스트 스트림 데이터를 시각화하는 방법을 제시하였다. Dynamic Maps는 동적으로 생성되는 지도로, Gansner는 트위터 데이터를 기반으로 동적으로 지도를 생성하는 방법을 제시하고 있다. 이를 통해 화두 되는 토픽이 무엇이고, 이를 다루는 저자들 간의 관계를 시간의 경과에 따라 지도

형식으로 시각화할 수 있다(Gansner et al., 2013).

시각화를 위한 프레임워크의 경우 Meyer는 데이터 시각화 구현에 있어서 데이터에 집중할 수 있는 시각화 프레임워크를 제안하였다. 기존에는 시각화 대상이 되는 데이터를 시각화 도구를 활용하여 구현하였다. 반면, Meyer의 시각화 프레임워크에서는 데이터 분석 도구에서 활용할 수 있도록 시각화 기법들을 일반화 시켰다. 이 연구는 데이터 분석 전문가들의 시각화 구현에 있어서 새로운 방법론은 제시하였다(Meyer et al., 2006).

이렇듯 텍스트 시각화에 있어서 효과, 분석 방법, 프레임워크 등 다양한 연구와 시도가 활발히 진행되어 왔다. 하지만, 텍스트 시각화를 위한 데이터 관점의 정보모델과 기능 관점의 공통 요구사항을 체계적으로 정리하고 있지는 못하다.

## 3. 텍스트 데이터 시각화 프레임워크

기존에 제시되었던 텍스트 데이터 시각화 프레임워크는 표준화 되어있지 못하였다. 특정 데이터에 의존적이거나 새로운 시각화 기법을 적용할 수 있도록 제안되었다. 따라서 이러한 프레임워크는 데이터나 시각화 방안이 변경되었을 때 활용될 수 없는 문제점이 존재하였다. 이에 따라 본 연구에서는 위와 같이 다양한 도메인의 데이터와 시각화 방안을 모두 적용할 수 있는 통합된 시각화 프레임워크를 제안한다.

제안하는 프레임워크는 유지보수성, 확장성 등이 우수한 MVC 디자인 패턴에 기반하여 설계한다. 즉, 시각화 방안이 표현되는 요소인 뷰를 제외한 모델과 컨트롤러만을 포함한 프레임워크를 제안함으로써 다양한 시각화 방안을 적용할 수 있는 프레임워크를 제안한다. 또한 최근 MVC 모델의 문제점을 해소하기 위해 등장한 MOVE 모델에서 역시 뷰를 제외한 모델, 오퍼레이션 그리고 이벤트 요소가 제안하는 프레임워크에 포함된다(Irwin, 2012).

다음 3.1절과 3.2절에서 도출된 두 가지 기준을 기반으로 프레임워크를 제시한다. 3.1절에서는 텍스트 데이터에 대한 보편적인 시각화 모델 제시한다. 그 후 3.2절에서는 현재 활용되고 있는 텍스트 데이터의 다양한 시각화 방안들을 분석하여 공통 요구사항을 도출한다. 우리는 이 두 가지를 만족시키는 프레임워크를 제안함으로써 다양한 텍스트 데이터를 모두 포함할 수 있는 시각화 프레임워크를 제안한다.

### 3.1. 텍스트 데이터 시각화 모델

텍스트 데이터는 대표적인 비정형 데이터이다. 특히 빅데이터(Big Data) 및 이와 관련한 기술들이 관심을 받으면서 비정형 데이터인 텍스트 데이터에 대한 관심도 높아지고 있다. 대표적인 텍스트 데이터로는 오피스 문서, 웹 상의 글과 댓글, 이메일, SMS, SNS 등 소셜 커뮤니케이션 데이터, 기타 이 외의 서지정보와 같은 다양한 메타데이터 상의 텍스트 정보들 등이 있다. 본 연구에서는 이러한 다양한 텍스트 데이터들을 대상으로 시각화 요소를 다음의 3가지의 관점에서 식별하고 이를 통합하여 시각화를 위한 공통된 정보모델을 도출한다.

• **구조적 관점:** 가장 기본적인 구성요소

를 구분하는 관점이다. 요소들은 텍스트 데이터가 갖고 있는 가장 보편적인 특성을 반영하여 그 모델을 구성하기 때문에 시각화 대상을 확인하는데 용이하다. 또한 문서 내 본문에서 단어 빈도수와 같이 메타데이터로 관리되는 형태적 특성 요소들 역시 중요한 구조적 관점의 요소이다.

• **상호작용 관점:** 문서 사이에서 발생하는 다양한 상호작용으로 구분하는 관점이다. 특히 대부분의 온라인문서는 상호작용을 고려하여 설계된다. 각각의 상호작용 요소는 문서 간의 연관관계나 문서 내에서 구성요소 간의 연관관계를 나타낸다. 이러한 연관관계의 시각화는 텍스트 데이터의 관계정보를 쉽게 이해할 수 있도록 돕는다.

• **내용적 관점:** 단어나 어구에 내용적 요소에 대한 태깅으로 표현된다. 본 연구에서는 현재 자연언어처리 기술로 자동으로 생성 가능한 수준의 태그들인 단어를 특정 개체와 연결시키는 개체명태그, 개체의 속성을 표현하는 속성태그, 개체들과 연루된 사건을 연결한 사건태그, 문장의 감성적 분위기를 표현하는 감성태그를 대상으로 한다.

〈Table 1〉 Visualization Elements for the Structural Characteristics of the Document

| Symbol | Description | Definition |
|---|---|---|
| $k_d$ | Keyword that appears in the Document $d$ | - |
| $TF_{kd}$ | Frequency of the Keyword $k_d$ | - |
| $DF_{ks}$ | Frequency of the Documents that include the Keyword $k$ in the Document Set $s$ | - |
| $w_{kd}$ | Weight of the Keyword $k_d$ | $w_{kd}$ = f($TF_{kd}$) |
| $F_d$ | Features of the Document $d$ | $F_d$ = Set of ($k_d$, $w_{kd}$) |

〈Table 2〉 Visualization Elements of the Document Interactions

| Symbol | Description | Definition |
|---|---|---|
| *SInfo* | Sender Information | $SInfo = (Id_{sender}, time_{sender})$ |
| *RInfo* | Receiver Information | $RInfo = (Id_{receiver}, Type_{receiver}, time_{receiver})$ |
| *DInteraction* | Interaction Information between the Documents | $DInteraction = (DID_{origin}, DID_{Related}, InteractionType)$ |

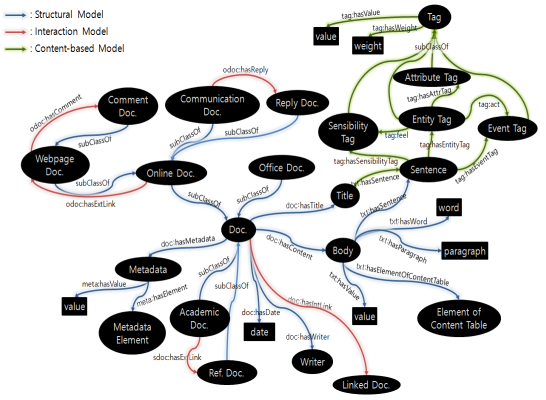
〈Table 3〉 Content-based Visualization Elements

| Symbol | Description | Definition |
|---|---|---|
| $NETag_d$ | Named Entity Tags that appear in the Document $d$ | $NETag_d$ = Set of ($Entity_{name}$, $Entity_{position}$, $ATag_e$) |
| $ATag_e$ | Attribute Tags of the Entity $e$ | $ATag_e$ = Set of ($Attr_{name}$, $Attr_{value}$) |
| $ETag_d$ | Event Tags that appear in the Document $d$ | $ETag_d$ = Set of ($Event_{name}$, $Event_{position}$) |
| $STag_d$ | Sensibility Tags that appear in the Document $d$ | $STag_d$ = Set of ($Sensibility_{type}$, $Sensibility_{score}$, $Sensibility_{feature}$) |
| $TagInfo_d$ | Combined Tag Information of Document d | $TagInfo_d$ = combine($NETag_d$, $ETag_d$, $STag_d$) |
| $DSim_{d1,d2}$ | Similarity between the Document $d1$ and $d2$ | $DSim_{d1,d2}$ = sim($Tag_{d1}$, $Tag_{d2}$) |
| $TopicRelation_{i,j}$ | Topic Relation Information between Keyword $i$ and $j$ | $TopicRelation_{i,j} = (i, j, Relation_{type}, Relation_{score})$ |

〈Table 4〉 Visualization Elements of the Integrated Model

| Symbol | Description | Definition |
|---|---|---|
| $ATagInfo_{x,y}$ | Combined Tag Information of the Document related the Author $x$ and $y$ | $ATagInfo_{x,y} = TagInfo(x) \wedge TagInfo(y)$ |
| $ARInfo_{x,y}$ | Relation Information between Document Author $x$ and $y$ | $ARInfo_{x,y}$ = (Set of ($ATagInfo_{x,y}$), $AuthorInteraction_{type}$, $AuthorInteraction_{score}$) |
| $Propagation_k$ | Propagated Tag Information of the Keyword $k$ | $Propagation_k$ = Set of ($\underline{DocumentInteraction_k}$, $TagInfo_d$) * Document $d$ has the Keyword $k$. |

위의 세 가지 관점에서의 표현된 정보모델을 도식화하면 다음의 <Figure 1>과 같은 통합 모델로 표현할 수 있다. 또한 다음 <Table 1, 2, 3, 4>는 이러한 각각의 관점과 통합적 관점에서 추가적으로 확인할 수 있는 시각화 요소에 대해 설명한다.




〈Figure 1〉 Integrated Model of the Text Data

## 3.2. 텍스트 데이터 시각화 도구의 공통 요구사항

텍스트 데이터 시각화는 데이터 시각화 연구의 한 분야로, 주로 텍스트의 형식, 구조, 내용, 패턴, 관계, 트렌드 등을 시각적으로 표현하여 텍스트가 내포하고 있는 정보를 보다 쉽게 파악하는 것을 목적으로 한다. 대표적으로 텍스트 내용 정보 전달을 목적으로 자주 사용되는 태그 클라우드, 텍스트 관계 정보 전달을 위해 사용되는 Email Map, 그리고 트렌드 정보 전달을 위한 토픽 랭크 등이 있다. 본 연구에서는 다양한 텍스트 데이터 시각화 기법들의 기능적 패턴을 정리하고, 텍스트 데이터 시각화를 위한 공통된 요구사항을 도출한다.

본 연구에서는 다양한 텍스트 데이터 시각화 기법을 목적 별로 분류하고, 각각의 기법에 대해 데이터 관점으로 요구사항을 정리하였다. 시각화 기법의 목적은 크게 텍스트 형식 및 구조 전달, 텍스트 내용 정보 전달, 텍스트 관계 정보 전달, 그리고 텍스트 트렌드 정보 전달로 나누었다. 데이터 요구사항은 텍스트와 관련된 정보로, 각 시각화 기법이 시각화를 위해 필요한 데이터에 해당한다. <Table 5>에서 11가지 시각화 기법들에 대한 데이터 요구사항을 정리하였다.

텍스트 구조 및 형식을 전달하기 위한 주요 시각화 기법으로는 가독성 정보 시각화와 변경 정보 시각화가 있다. 가독성 정보 시각화는 문장의 길이에 비례하여 텍스트의 명도가 낮아지도록 표현하여 문서를 직접 읽지 않고도 책의 가독성 정보를 직관적으로 파악할 수 있도록 표현한 방법이다(Kim and Park, 2013). 이는 문장의 길이가 길수록 가독성이 떨어진다는 연구

〈Table 5〉 Data Requirements and Usages of the Visualization Technique

| Category | Visualization Technique | Required Data | What to Know |
|---|---|---|---|
| Format & Structure Information | Readability Information | - Sentences of the Document | - Readability Info. of the Documnet |
| | Change Information | - Sentences of the 2 Documents<br>- List of Change Info.(pos & string) | - Change Info. between the Same Documents |
| Content Information | Tag Cloud | - Keywords(word & freq.) of the Document | - Important Keywords of the Document |
| | Document Bubble Chart | - List of the Document Info. (title, type & size) | - Relative Document Size |
| | TextArc | - List of the Sentence Info. (sentence & words)<br>- Keywords(word & freq.) of the Document<br>- List of the Keyword Relation of the Document(2 words & weight) | - Keywords Information of the Document<br>- Keyword Relation Information |
| | Chord Diagram with Split Links (Keyword Info.) | - List of the Document Info. (title, type & keywords)<br>* keywords = (word & freq.) List | - Important Keywords of the Document<br>- Common Keywords Information<br>- Relative Keyword Importance |
| | Tree Map (Document Info.) | - List of the Document Info. (title, cited count & type)<br>- Document Type Hierarchy Info. | - Important Document<br>- Hierarchy of the Document<br>- Relative Document |
| Relation Information | Email Map | - List of the Communication Document Relation (from, to, cc & bcc) | - Close People Relation Info.<br>- Inferred Important People |
| | Author Relation Network | - List of the Author Info. (name & weight)<br>- List of the Author Relation (2 authors & weight) | - Close People Relation Info.<br>- Important People<br>- Inferred People Relation |
| Trend Information | Keyword Trend Timeline | - Periodical Freq. Info. List of the Keyword(period, word & freq.) | - Keyword Popularity Trend<br>- Inferred Time of the Keyword-related Issue |
| | Topic Rank Round Tree | - Related Keywords Info. (word, related words) | - Important Related Keyword |

결과를 바탕으로 구현된 시각화 기법이다. 다음의 변경 정보 시각화는 두 문서를 비교하여 동일한 문장일 경우와 추가, 삭제, 또는 변경된 문장인 경우를 구분하여 서로 다른 색상으로 표현하는 방법이다. 두 문서 사이에서 변경된 문장의 비중을 직관적으로 파악할 수 있고, 문서가 개정되었을 때 두 문서의 비교에 유용하다(Posavec and McInerny, 2009).

텍스트 내용 정보를 전달하기 위한 시각화하기 위한 기법에는 다양한 방법이 존재한다. 그 중에서도 태그 클라우드는 대표적인 문서의 중요 키워드 정보의 시각화 기법이다. 문서의 내용을 분석해서 도출된 키워드들을 그 중요도에 따라 크기를 다르게 하여 표현한다(Halvey and Keane, 2007). 이 표현 방법은 문서의 주요 내용과 정보 간 상대적 중요도를 한 눈에 파악

할 수 있도록 돕는다. 두 번째로 소개할 시각화 방안인 Document Bubble Chart는 문서에서 사용된 전체 단어의 수에 비례한 크기의 원으로 표현하고 문서의 타입에 따라 색을 구분한다(Yau, 2010). 문서마다의 내용량을 비교할 수 있고 타입 별로 대략적인 문서의 개수와 규모를 한눈에 확인할 수 있다. 다음은 TextArc라는 시각화 기법이다. 문서 내의 중요 단어를 파악하고 단어 간의 관계를 비교하는데 유용한 시각화 방안으로 타원의 가장자리에 매우 작은 폰트로 이루어진 문장들을 나열하여 선으로 표현하고, 보이지 않는 스프링으로 연결된 단어들을 나열하며, 이때 문서에 많이 등장하는 단어는 크게 표현이 하고, 드물게 사용되는 단어들은 같은 공간을 공유하여 나열한다(Paley, 2009). 이를 활용하면 텍스트 내의 단어의 중요성과 단어 간의 관계를 쉽게 파악할 수 있다. Chord Diagram은 노드 간의 상호작용 관계를 링크로 직관적으로 표현하기 위해 제시된 시각화 방안이다(Hall, 2013). 그러나 본 연구에서 소개하는 Chord Diagram with Split Links는 링크의 분리를 허용함으로써 상호작용이 아닌 문서마다의 공통요소의 규모를 비교하기 위한 시각화 방안이다. 즉 텍스트 데이터에서는 문서를 노드, 문서 간의 공통 키워드를 링크로 표현한다. 노드들은 원형의 경로를 따라 나열되어 고리를 구성하고, 링크는 고리 내부에 곡선의 링크로 표현한다. 노드의 길이는 텍스트에 포함된 모든 키워드의 가중치의 합에 의해 결정되고, 노드의 색은 문서의 타입, 링크의 두께는 문서에서 키워드의 빈도에 따라 표현된다. 이렇게 시각화된 다이어그램을 통해 키워드가 문서의 종류에 따라 얼마나 중요하게 사용되고 있고 얼마나 많은 문서에서 사용되고 있는지 한눈에 파악할 수 있고 문서 내에서 키워드의 비중을 파악할 수도 있다. 마지막으로 Tree Map은 직사각형의 맵으로 크기와 계층구조를 갖는 객체를 시각화하는 방법이다(Anton, 2013). 문서 정보를 표현하기 위해 사용되는 Tree Map의 경우, 각 문서의 가장 상위 단계의 분류는 하위 분류의 가중치의 합에 의해 결정되어 직사각형의 섹션을 나눈다. 이 때 문서의 가중치는 설정하기에 따라 다르나 문서의 피인용 횟수와 같이 정량적인 성질이 존재하는 값으로 설정하여 그 크기에 의미를 부여한다. 이와 같은 섹션은 문서로 분리될 때까지 직사각형으로 나누어진다. 또한 트리는 키워드의 가장 상위 분류에 따라 색을 구별하여 표현한다. 이는 문서 및 문서 집합이 차지하는 상대적 비중을 직관적인 확인을 돕고 문서의 계층관계 역시 파악할 수 있다.

텍스트 관계 정보 전달을 위해서도 다양한 시각화 방안이 존재한다. Email Map은 메일을 보낸 사람(from), 받는 사람(to), 참조(cc, bcc)로 구분하여 주소록 내의 각 개인 간 관계 정보를 시각화하여 데이터 내부에 가려진 관계 정보를 직관적으로 보여준다(Baker, 2007). 사람이 노드, 사람 간에 전송된 메일이 방향성이 존재하는 링크로 표현된다. 다음으로 소개할 저자 관계 네트워크 시각화는 문서 간의 관계를 넘어 문서의 저자에 대한 정보를 표현하는 방안이다(Polley, 2013). 주요 저자의 노드를 크게 표시하고 저자 간의 상호작용(인용 텍스트의 저자/공동 저자)이 많은 저자들에 대해서 링크의 명도를 진하게 표시한다. 이를 통해 주요 저자와 저자 간의 관계를 파악할 수 있다. 예를 들어 문서 작성 활동에 많이 참여하고 저자의 문서가 많은 인용을 받은 경우 주요 저자라고 판단

할 수 있다. 또한 많은 수의 저자와 공동 문서 작성에 참여한 경우 교수와 학생의 관계임을 유추할 수도 있다.

텍스트 트렌드 정보를 전달하기 위한 시각화 방안에는 키워드 트렌드 타임라인과 토픽 랭크 원형 트리가 있다. 키워드 트렌드 타임라인은 2차원의 그래프 형태로 시각화된다(Lars, 2011). x축으로 시간의 흐름을 표현하고 키워드의 출현 빈도는 y축의 값으로 표현된다. 이를 통해 기간별 키워드의 트렌드를 분석할 수 있고, 특정 시점에 눈에 띄게 큰 값이 나타날 경우 키워드에 관한 특정 사건이 발생했음을 유추할 수 있다. 토픽 랭크는 키워드에 대한 연관 키워드 중 연관성이 높은 키워드를 분석하여 순차적으로 표현한 목록으로 각 연관 키워드는 하위 연관 키워드를 지닐 수 있다. 토픽 랭크 원형 트리는 위의 결과 데이터를 방사형으로 표현하고 랭킹이 높을수록 노드의 색을 진하게 표현하여 시각화 함으로써 주요 연관 키워드를 용이하게 파악할 수 있도록 돕는다(Saltlux, 2014).

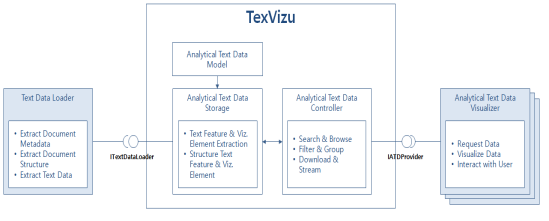
### 3.3. Tex Vizu (Framework of the Text Data Visualization)

본 절에서는 앞서 도출된 텍스트 데이터 시각화 모델과 시각화 도구의 공통된 요구사항을 바탕으로 텍스트 데이터 시각화를 위한 프레임워크인 TexVizu를 제시한다. <Figure 2>에 표현된 것과 같이 TexVizu는 수집된 텍스트 데이터를 분석하여 시각화를 위한 공통된 분석적 데이터 모델에 기반한 저장 및 관리구조를 제공하며, 텍스트에 대한 통일된 조회・검색・전송 체계를 제공함으로써 다양한 시각화 도구에 적용 가능한 기반을 제공하는 프레임워크이다.

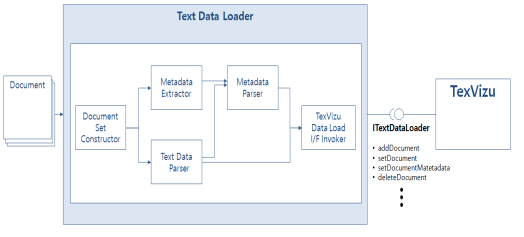
프레임워크는 크게 분석적 텍스트 데이터 모델(Analytical Text Data Model, ATDM), 텍스트 데이터 적재기(Text Data Loader, TDL), 분석적 텍스트 데이터 저장소(Analytical Text Data Storage, ATDS), 분석적 텍스트 데이터 제어기(Analytical Text Data Controller, ATDC), 분석적 텍스트 데이터 시각화 도구(Analytical Text Data Visualizer, ATDV)와 같은 5가지 구성요소를 갖는다.

ATDM은 3.1절에서 정의한 텍스트 데이터의 시각화 요소들을 구조화한 데이터 모델로서 분석적 텍스트 데이터의 저장, 관리, 활용에 기반이 된다.

TDL은 수집된 텍스트 데이터를 ATDS에 저장하는 역할을 담당하는 가상의 구성요소로서 TexVizu가 제공하는 적재 인터페이스인




〈Figure 2〉 Framework of the Text Data Visualization




〈Figure 3〉 Text Data Loader (TDL) Internal Process

ITextDataLoader를 준수하여 작동되어야 한다. 통상 텍스트 데이터는 기관 내의 레거시 시스템이나 웹 상의 서비스 등 다양한 출처로부터 수집되고, 수집된 형상도 다양하다. 그러므로 수집된 텍스트 데이터를 통일된 형식으로 변환하여 적재해야 한다. 이때 TDL은 수집된 문서의 메타데이터와 문서 내의 순수한 텍스트 데이터를 추출하여 ATDS에 전달해야 한다. <Figure 3>은 이와 같은 TDL의 내부 동작 구조를 표현한 구성도이다.

〈Table 6〉 TDL pseudo code

```
Procedure TDL
if getDocumentSetInfo(docSet.id) is null
      call addDocumentSet(docSet.name, docSet.type,
      docSet.id)
end

extract documents from docSet
for i=0 to documents.size do
      extract metadata from documents(i)
      standardize metadata
      call addDocument(docSet.id, documents(i).text,
      metadata)
end
```

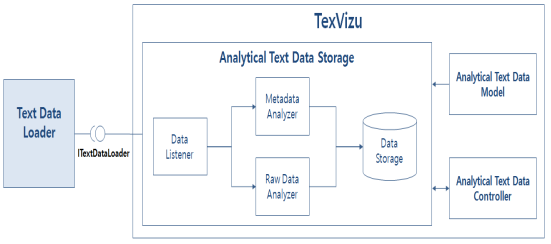
<Table 6>의 TDL 수도 코드(pseudo code)에서 확인 가능하듯이 ITextDataLoader (ITDL)는 TexVizu에 데이터를 추가하기 위하여 몇 가지 인터페이스가 사용된다. TDL은 문서를 기록하기 전에 'getDocumentSetInfo'를 호출하여 현재 기록하려는 문서셋에 대한 정보가 TexVizu에 기록이 되어있는지 확인하고 기록되어 있지 않다면 'addDocumentSet' 인터페이스를 호출하여 추가한다. 그 후 추가하려는 문서 각각에 대하여 메타데이터를 추출하고 xml, json 등 표준화된 포멧으로 설정하고 이를 기록하기 위한 인터페이스로 'addDocument'를 호출한다. ITDL은 데이터 추가를 위한 용도 이외에도 관리를 위한 인터페이스가 포함되어 있고 <Table 7>에서 설명한다.

<Figure 4>의 ATDS는 TDL이 적재한 텍스트 데이터를 ATDM에 따라 분석구조화하여 저장한다. 이때, ATDS는 비정형 데이터의 텍스트 데이터로부터 시각화 요소를 추출하여

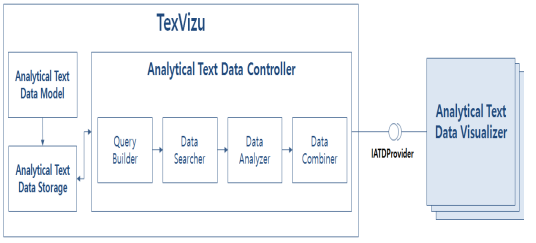
〈Table 7〉 ITDL Interface

| Return Type | Interface Name | Parameter | Description |
|---|---|---|---|
| void | addDocumentSet | String docSet.name<br>String docSet.type<br>String docSet.id | - Add the Document Set |
| DocumentSet | getDocumentSetInfo | String docSet.id | - Get the Document Set Metadata |
| void | deleteDocumentSet | String docSet.id | - Delete the Document Set |
| void | addDocument | String docSet.id<br>String text<br>String metadata | - Add the Document |
| void | setDocument | String docSet.id<br>String doc.id<br>String text | - (Re)Set the Text Data |
| void | setMetadata | String docSet.id<br>String doc.id<br>String metadata | - (Re)Set the Metadata |
| void | deleteDocument | String docSet.id<br>String doc.id | - Delete the Document |

정형화하는 역할을 담당한다. 이렇게 함으로써 함께 제공된 메타데이터와 함께 추출되어 정형화된 시각화 요소들은 프레임워크가 시각화 도구에게 제공하는 데이터의 근간을 이루게 된다.




〈Figure 4〉 Analytical Text Data Storage (ATDS) Internal Process




〈Figure 5〉 Analytical Text Data Controller (ATDC) Internal Process

〈Table 8〉 ATDS pseudo code

```
Procedure ATDS
if data.type is Metadata
    extract metadata from data
    for i=0 to metadata.size
        analyzed_data = analyze_metadata(metadata(i))
        add analyzed_data into result_metadata
    end
    store result_metadata into DB.meta_table
end
if data.type is Text
    while analysis_list is scanned
        analyzed_data = analyze_text(data.data, data.
        element, analysis.type)
        store analyzed_data into DB.analysis_table
    end
    store data into DB.text_table
end
```

<Figure 5>의 ATDC는 ATDS에 대한 표준화된 활용체계를 제공한다. 적재되어 분석된 텍스트 데이터를 활용하기 위해서는 시각화 목적에 알맞은 텍스트 데이터를 검색하거나 탐색하여 선택할 수 있어야 한다. 또한, 선택된 데이터에 포함되어 있는 불필요한 데이터들을 걸러내어 정제하거나, 비슷한 텍스트 데이터를 군집화하여 시각화의 범위를 축소하는 등 데이터 수준에서의 제어 기능을 제공해야 한다.

ATDV는 ATDC가 제공하는 인터페이스인 IATDProvider를 통해 분석적 텍스트 데이터를 공급받는 가상의 시각화 도구이다. IATDProvider가 제공하는 메쏘드는 <Table 10>에서 설명한다. 메쏘드는 ATDS에 저장된 정보를 단순히 조회하는 조회 메쏘드(getDocumentSet, searchDocuments 등)와 호출 시 분석을 수행하는 분석 메쏘드(getTopics, searchPeople 등)로 구분된다. IATDProvider는 이러한 메쏘드를 이용하여 TexVisu와 ATDV의 독립성을 유지한다.

〈Table 9〉 ATDC pseudo code

```
Procedure ATDC
Build query
Load data from DB using query
while analysis_list is scanned
    analyzed_data = analyze_data(data, analysis.type)
    refine analyzed_data
    add analyzed_data to data_list
end
result_data = combined_data_list(data_list, analysis.type)
return result_data
```

<Table 10> IATDProvider Interface

| Return Type | Interface Name | Parameter | Description |
|---|---|---|---|
| DocumentSet | getDocumentSet | Stirng docSet.id | - Get the Document Set Metadata |
| DocSearchResult | searchDocuments | Stirng docSet.id<br>String query<br>int pageSize = 0<br>int pageNo = 0 | - Search Documents from Query<br>- (optional) Set the Page Size & No. |
| int | countDocuments | String docSet.id<br>String query | - Get the total Document Count |
| String | getDocumentMetadata | String doc.id<br>String key = null | - Get the '*key*' Metadata<br>(if *key* is null, get entire Metadata) |
| KeywordSearchResult | getTopics | String docSet.id<br>String query<br>int pageSize = 0<br>int pageNo = 0 | - Get the Query-related Topics<br>- (optional) Set the Page Size & No. |
| PersonSearchResult | searchPeople | String query<br>int pageSize = 0<br>int pageNo = 0 | - Search People from Query<br>- (optional) Set the Page Size & No. |

## 4. 사례연구

사례연구에서는 이미 개발되어 서비스된 실제 시각화 사례에 대해서 앞서 제시한 TexVizu의 인터페이스인 IATDProvider를 구현가능성 여부를 검증함으로써 보편성을 증명한다. 첫 번째 사례로서는 소셜 빅데이터를 이용하여 정치인들의 여론 피드백을 분석한 "트루스토리 시즌1 정치인"의 시각화를 채택하였다. 두 번째 사례로서는 온라인 상의 뉴스를 수집하여 분석한 "O2D2 서비스"의 시각화를 채택하였다. 각각의 사례는 피드백 미디어로서의 소셜 빅데이터와 푸쉬 미디어로서의 온라인 뉴스를 대상으로 하고 있으며, 본 연구를 포함한 시각화 연구의 과정에서 개발된 시연용 서비스이다.

### 4.1. 정치인 여론 피드백 시각화 사례

트루스토리는 소셜 데이터에서 하나의 토픽을 선정하여 그에 대한 다양한 분석 결과를 시각화하여 웹에 표현하는 서비스이다. 트루스토리에서는 소셜 감성 분석, 연관 키워드 분석, 그리고 트렌드 분석을 수행하고 기간 별로 분석되어 분석 결과의 변화를 확인 할 수 있다.

〈Figure 6〉 Social Sensibility Analysis : 'TrueStory - Politician'

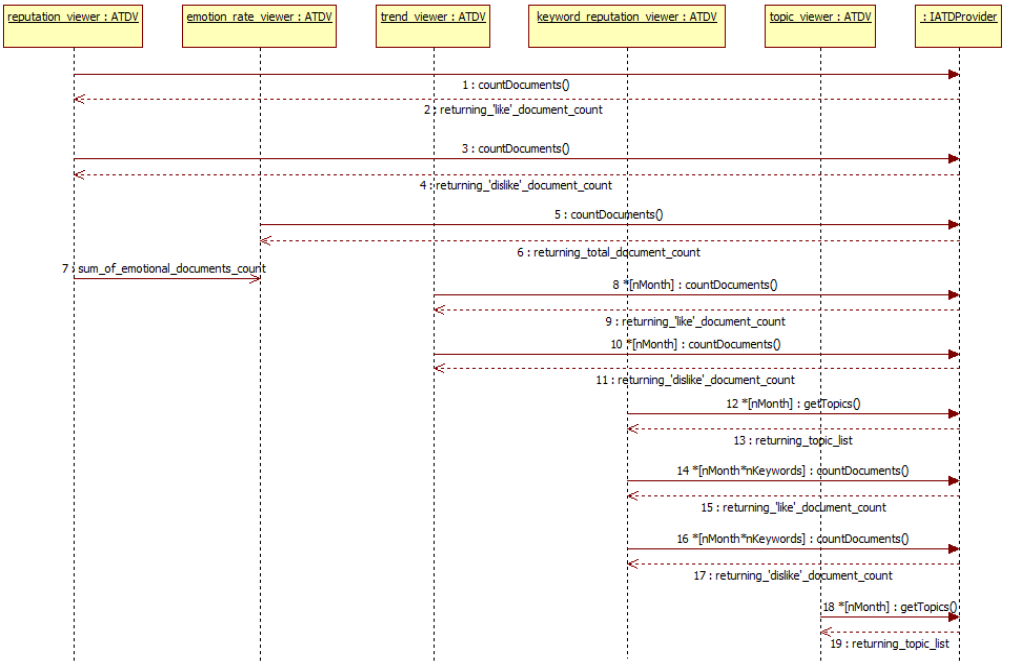
소셜 감성 분석은 소셜 데이터에 나타난 토픽에 대한 호감 정보를 분석하여 시각화하는 서비스로 세부 분석 항목으로는 통합 호감 지

수, 감성 비율, 월별 호감 트렌드, 함께 언급된 키워드에 따른 호감 지수가 있다. <Figure 5>에 나타난 분석 시각화 항목들은 1년 간의 소셜 데이터에 기록된 결과를 반영한 것으로 각각은 countDocuments 메쏘드의 쿼리를 다르게 설정하여 호출함으로써 그 결과 데이터를 얻을 수 있다. 즉, 쿼리는 정치인에 관한 트위터 문서들 중 호감 정보를 담고 있는 메타데이터 항목을 확인함으로써 문서들을 분류할 수 있도록 구성하고 그 결과 값을 시각화 한다. 함께 언급된 키워드에 따른 호감 지수는 getTopics를 통해 정치인에 관한 주요 연관 키워드를 얻고 연관 키워드까지 포함하여 countDocuments의 쿼리를 구성할 경우 시각화를 위한 결과 값을 얻을 수 있다.


〈Figure 7〉 Related Keyword Analysis : 'TrueStory - Politician'

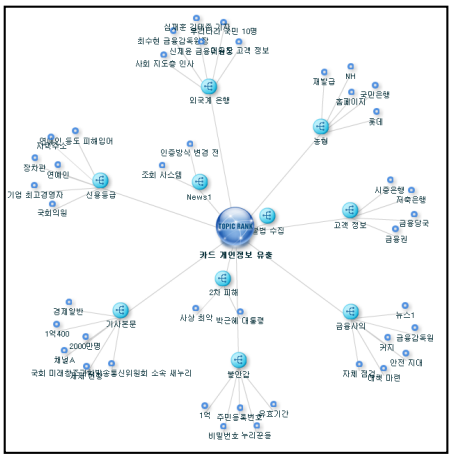
<Figure 7>의 연관 키워드 분석은 소셜 데이터에 토픽 키워드에 대한 월별 연관 키워드를 중요도 순으로 보여주는 것이다. 해당 분석은 getTopics 인터페이스 메쏘드를 월별로 호출하여 월별 주요 연관 키워드를 추출한다.

트루스토리 서비스의 시각화 항목들을 시퀀스 다이어그램으로 표현하면 <Figure 8>과 같




〈Figure 8〉 Sequence Diagram : 'TrueStory-Politicain'


〈Figure 9〉 'O2D2' Service Main Page




〈Figure 10〉 Topic Rank of 'O2D2'

다. 제안된 IATDProvider 인터페이스로 동작이 가능함을 확인 할 수 있다.

### 4.2. 뉴스 분석 및 검색 시각화 사례

<Figure 9>는 웹 뉴스 데이터를 분석하여 정보를 제공하는 서비스인 'O2D2'의 웹 페이지 화면이다. 그림에서 확인 가능하듯이 지정 날짜에 대한 뉴스 정보를 제공한다. 첫 화면의 메인을 장식하는 투데이 이슈 시각화 서비스는 해당 날짜 주요 뉴스의 랭킹을 매겨 원의 둘레를 따라 시각화한다. 이는 searchDocuments를 통해 주요 뉴스 문서를 추출하게 되고 각각의 뉴스 문서는 getDocumentMetadata 인터페이스를 호출하여 뉴스 문서에 포함된 키워드의 목록을 추출하여 얻은 결과를 통합하여 시각화한다. 또한 선택된 뉴스 문서의 연관 토픽 항목은 getTopics를 호출하여 시각화 한다. 이 때 쿼리는 앞서 추출한 키워드를 기반으로 구구성한다. 트랜드 차트는 키워드에 대한 월별 생성된 문서의 양을 그래프 형태로 시각화한 항목으로 countDocuments를 사용하여 결과 값을 얻을 수 있다.

<Figure 10>은 'O2D2' 서비스에서 시맨틱 검색을 수행했을 때 나타나는 결과 중 하나인 토픽 랭크의 시각화 결과로 토픽 랭크 원형 트리 시각화 도구를 사용하였다. 해당 시각화를 수행하기 위해서 시맨틱 검색 시 사용한 키워드를 입력 쿼리로 하는 getTopics를 호출한다. 해당 분석을 통해 나온 연관 키워드를 쿼리에 포함하여 getTopics를 다시 호출함으로써 주제 키워드와 1-depth 연관 키워드를 모두 포함한 2-depth의 연관 키워드를 획득할 수 있고 이를 다시 원형 트리 형태로 시각화한다.

이러한 'O2D2' 서비스의 시각화 항목들의 수행 흐름을 시퀀스 다이어그램으로 표현하면 <Figure 11>와 같다. 각각의 시각화 항목들은 쿼리의 입력 값만 주어진다면 독립적으로 시각화 가능한 형태를 지니고 있고 제안된 인터페이스로 동작이 가능함이 확인 가능하다.


〈Figure 11〉 Sequence Diagram : 'O2D2' Service

### 4.3. 사례연구 시사점

<Table 11>는 각각의 서비스 시각화에 필요한 인터페이스를 체크한 표이다. 현재 제공된 서비스들은 하나의 문서셋에 대해서 서비스가 제공되는 형태이기 때문에 메쏘드 getDocumentSet과 searchPeople은 활용되지 않지만 다양한 문서셋이 결합되어 사용되는 경우 나 저자 정보에 대한 시각화를 수행하는 경우 활용될 수 있다. 또한 위의 표에서 확인 가능하듯이 사례에서 제공되는 서비스는 제안된 인터페이스를 통해 모두 시각화가 가능하다. 이는 제안된 프레임워크가 텍스트 데이터 시각화를 위한 다양한 형태의 데이터를 제공할 수 있음을 증명한다.

## 5. 결론

본 연구에서는 텍스트 데이터의 시각화에서 재사용성, 유지보수성 및 확장성을 확보하기

〈Table 11〉 Visualization Interface Check List of the Case Studies

| Visualization Interface | TrueStory - Politician | | | | | O2D2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Integrated Feeling Score | Sensibility Rate | Monthly Feeling Trend | Related Feeling Score | Monthly Related Keyword | Today Issue | Related Topic | Trend Chart | Topic Rank |
| getDocumentSet | | | | | | | | | |
| searchDocuments | | | | | | O | | | |
| countDocuments | O | O | O | O | | | | O | |
| getDocumentMetadata | | | | | | O | | | |
| getTopics | | | | O | O | | O | | O |
| searchPeople | | | | | | | | | |

위한 방안으로 MVC 디자인 패턴을 적용한 텍스트 데이터 시각화 프레임워크를 제시하였다. 텍스트 데이터 시각화 프레임워크 TexVizu는 텍스트 데이터로부터 분석되어 도출될 수 있는 텍스트 데이터 시각화 모델과 다양한 텍스트 시각화 기법들이 요구하는 데이터 요구사항을 바탕으로 구성되어 있다. TexVizu는 분석적 텍스트 데이터 모델이 기반하여 분석적 텍스트 데이터 저장소를 유지관리한다. 분석적 텍스트 데이터는 저장소는 ITextDataLoader 인터페이스를 지원하는 다양한 텍스트 데이터 적재기와 연동할 수 있다. 분석적 텍스트 데이터 컨트롤러를 통해 다양한 텍스트 시각화 도구에 텍스트 데이터 시각화를 위해 표준화된 인터페이스인 IATDProvider를 제공한다. 이처럼 TexVizu는 단지 텍스트 시각화 도구와의 결합도 만이 아니라 텍스트 소스와 결합도도 최소화할 수 있는 구조를 가지고 있다.

한편 텍스트 데이터로부터 추출된 정보들 중 개체에 대한 데이터(사람, 조직, 지역, 건물, 사건 등)들은 각각의 고유의 속성을 같은 정보 구성요소이다. 그러므로 향후 연구에서는 TexVizu를 이러한 객체 데이터들을 저장, 관리하고 다양한 객체 정보의 활용을 지원할 수 있는 형태로 발전시키고자 한다.

## 참고문헌 (References)


Heijs, A., *Big Data: Rethinking Text Visualization*, Treparel, 2013. Available at http://treparel.com/wp-content/uploads/2012/07/WP-Big-Data-Rethinking-Text-Visualization.pdf(Downloaded 5 February 2014).

Baker, C., *Email Map*, Christopher Baker 2004-2014, 2007. Available at http://christopherbaker.net/projects/mymap ( Accessed 5 February 2014).

Gamma, E., R. Helm, R. Johnson, and J. Vlissides, *Design patterns: elements of reusable object-oriented software*, Addison-Wesley, Massachusetts, 1995.

Gansner, E., Y. Hu, and S. North, "Interactive Visualization of Streaming Text Data with Dynamic Maps," *Journal of Graph Algorithms and Applications*, Vol.17, No.4(2013), 515~540.

Hall S., *Chord Diagrams in D3*, 2013. Available at http://www.delimited.io/blog/2013/12/8/chord-diagrams-in-d3(Accessed 15 April 2014).

Halvey M., and M. T. Keane, "An Assessment of Tag Presentation Techniques," *Proceedings of the 16th international conference on World Wide Web,* (2007), 1313~1314.

Irwin, C., *MVC is dead, it's time to MOVE on*, 2012. Available at http://cirw.in/blog/time-to-move-on (Accessed 15 April 2014).

Jahiruddin, M. Abulaisa, L. Dey, "A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora," *Journal of Biomedical Informatics*, Vol.43, No.6 (2010), 1020~1035.

Jones, S., S. Payne, B. Hicks, and L. Watts, "Visualization of Heterogeneous Text Data in Collaborative Engineering Projects," *The 3rd IEEE Workshop on Interactive Visual Text Analytics,* (2013).

Kim, H. Y. and J. W. Park, *"A Review on Expressive Materials and Approaches to Text Visualization," Journal of Korea Contents Association*, Vol.13, No.1(2013), 64~72.

Kim, H. Y. and J. W. Park, “Case Analysis of Bible Visualization based on Text Data Traits - Focused on Content, Structure, Quotation of Text,” *Journal of Korea Contents Association*, Vol.13, No.8 (2013), 83~92.

Lars, *Essential Tools for Keyword Trend Analysis,* 2011. Available at http://www.tripwiremagazine.com/2011/07/keyword-trend-analysis.html(Accessed 5 February 2014).

Lee, J. Y, “A Usability Evaluation on the Visualization of Information Extraction Output,” *Journal of the Korean Society for Library and Information Science,* Vol.39, No.2(2005), 287~304.

Meyer, M., T. Gîrba, M. Lungu, “Mondrian: an agile information visualization framework,” *SoftVis '06 Proceedings of the 2006 ACM symposium on Software visualization,* (2006), 135-144.

Paley, W. B., *TextArc: Alice's Adventures in Wonderland,* 2009. Available at http://www.textarc.org/(Accessed 5 February 2014).

Polley T., *Studying Four Major NetSci Researchers (ISI Data),* 2013. Available at http://wiki.cns.iu.edu/pages/viewpage.action?pageId=2200066(Accessed 15 April 2014).

Posavec, S., G. McInerny, *The Evolution of the Origin of Species*, 2009. Available at http://www.itsbeenreal.co.uk/index.php?/on-going/about/(Accessed 5 February 2014).

Saltlux, *TopicRank*, 2014. Available at http://www.saltlux.com/topicrank/(Accessed 15 April 2014).

Yau, N., *How to Make Bubble Charts*, 2010. Available at http://flowingdata.com/2010/11/23/how-to-make-bubble-charts/(Accessed 15 April 2014).

Zhao, Q., P. Mitra, "Event Detection and Visualization for Social Text Streams," *International Conference on Weblogs and Social Media,* (2007).

Abstract

# A MVC Framework for Visualizing Text Data

Kwang Sun Choi* · Kyo Sung Jeong* · Soo Dong Kim**

As the importance of big data and related technologies continues to grow in the industry, it has become highlighted to visualize results of processing and analyzing big data. Visualization of data delivers people effectiveness and clarity for understanding the result of analyzing. By the way, visualization has a role as the GUI (Graphical User Interface) that supports communications between people and analysis systems. Usually to make development and maintenance easier, these GUI parts should be loosely coupled from the parts of processing and analyzing data. And also to implement a loosely coupled architecture, it is necessary to adopt design patterns such as MVC (Model-View-Controller) which is designed for minimizing coupling between UI part and data processing part. On the other hand, big data can be classified as structured data and unstructured data. The visualization of structured data is relatively easy to unstructured data. For all that, as it has been spread out that the people utilize and analyze unstructured data, they usually develop the visualization system only for each project to overcome the limitation traditional visualization system for structured data. Furthermore, for text data which covers a huge part of unstructured data, visualization of data is more difficult. It results from the complexity of technology for analyzing text data as like linguistic analysis, text mining, social network analysis, and so on. And also those technologies are not standardized. This situation makes it more difficult to reuse the visualization system of a project to other projects. We assume that the reason is lack of commonality design of visualization system considering to expanse it to other system. In our research, we suggest a common information model for visualizing text data and propose a comprehensive and reusable framework, TexVizu, for visualizing text data. At first, we survey representative researches in text visualization era. And also we identify common elements for text visualization and common patterns among various cases of its. And then we review and analyze elements and patterns with three different viewpoints as structural viewpoint, interactive viewpoint, and semantic viewpoint. And then we design an integrated model of text data which represent elements for visualization. The structural viewpoint is for identifying structural element from

* Strategic Business Center, Saltlux
** Corresponding Author: Soo Dong Kim
Department of Computer Science, Soongsil University
Soongsil University 369 Sando-Ro, Dongjak-Gu, Seoul, Korea
Tel: +82-2-824-0909, Fax: +82-02-815-0518, E-mail: sdkim777@gmail.com

various text documents as like title, author, body, and so on. The interactive viewpoint is for identifying the types of relations and interactions between text documents as like post, comment, reply and so on. The semantic viewpoint is for identifying semantic elements which extracted from analyzing text data linguistically and are represented as tags for classifying types of entity as like people, place or location, time, event and so on. After then we extract and choose common requirements for visualizing text data. The requirements are categorized as four types which are structure information, content information, relation information, trend information. Each type of requirements comprised with required visualization techniques, data and goal (what to know). These requirements are common and key requirement for design a framework which keep that a visualization system are loosely coupled from data processing or analyzing system. Finally we designed a common text visualization framework, TexVizu which is reusable and expansible for various visualization projects by collaborating with various Text Data Loader and Analytical Text Data Visualizer via common interfaces as like ITextDataLoader and IATDProvider. And also TexVisu is comprised with Analytical Text Data Model, Analytical Text Data Storage and Analytical Text Data Controller. In this framework, external components are the specifications of required interfaces for collaborating with this framework. As an experiment, we also adopt this framework into two text visualization systems as like a social opinion mining system and an online news analysis system.

**Key Words** : Text Data, Visualization, MVC Framework

Received: March 27, 2014 Revised: April 28, 2014 Accepted: May 12, 2014

# 저 자 소 개

**최 광 선**
소속 : 숭실대학교 컴퓨터학과 박사과정 (㈜솔트룩스 전략사업본부 본부장)
e-mail : kschoi@saltlux.com
2005년 숭실대학교 소프트웨어 공학 (석사)
2008년 숭실대학교 컴퓨터학 (박사수료)
1995년 ~ 2000년 대우정보시스템 기술연구소 대리
2000년 ~ 2002년 에이전트리더 기술연구소 실장
2003년 ~ 2005년 큐브테크 기술연구소 차장
2005년 ~ 현재 솔트룩스 전략사업본부 본부장
관심 분야 : 서비스 지향 아키텍처(SOA), 모바일 서비스(Mobile Service), 객체지향 S/W 공학, 컴포넌트 기반 개발(CBD), 소프트웨어 아키텍처(Software Architecture), 데이터 마이닝(Data Mining), 지식공학 (Knowledge Engineering)

**정 교 성**
소속 : ㈜솔트룩스 전략사업본부 사원
e-mail : ksjeong@saltlux.com
2013년 한양대학교 전자컴퓨터통신공학과 (석사)
2013년 ~ 현재 솔트룩스 전략사업본부 사원
관심 분야 : 데이터 마이닝(Data Mining), 데이터베이스(Database), 지식공학(Knowledge Engineering)

**김 수 동**
e-mail : sdkim777@gmail.com
1984년 Northeast Missouri State University 전산학 (학사)
1988년/1991년 The University of Iowa 전산학 (석사/박사)
1991년~1993년 한국통신 연구개발단 선임연구원.
1994년~1995년 현대전자 소프트웨어연구소 책임연구원.
1995년 9월~현재 숭실대학교 컴퓨터학부 교수.
관심분야 : 서비스 지향 아키텍처(SOA), 클라우드 컴퓨팅(Cloud Computing), 모바일 서비스(Mobile Service), 객체지향 S/W공학, 컴포넌트 기반 개발(CBD), 소프트웨어 아키텍처(Software Architecture)