웹 사이트의 페이지 연관성에 관한 시각화 연구

A Study of Visualization by Page Connection of Web Sites

김영진, Youngjin Kim*, 이경원, Kyungwon Lee*
*아주대학교 미디어학과

요약 인터넷 웹사이트의 규모가 커지면서 그 안에 담고 있는 정보의 양과 종류가 많아지고 복잡해 지고 있다. 이에 사용자는 같은 사이트 내에서도 정보를 찾는 것에 어려움을 느끼고 있다. 이에 대한 해결책으로써 웹사이트 안에 있는 각 페이지들의 하이퍼링크 정보로부터 얻어낸 연결 정보를 분석하여 그 안에서 관계성을 추출 하고 이를 관련 있는 페이지들끼리의 모임으로 분류해서 시각화 하는 방법을 제안하였다. 본 논문에서는 시각화 인터페이스로써 태양계의 행성들을 메타포로 이용하였다. 즉 웹사이트 안의 페이지는 하나의 행성의 모습으로 표현되고, 페이지들의 하이퍼링크에 의한 연결된 수는 중력으로써 다른 페이지를 끌어 당기는 힘으로 사용된다. 이때 행성의 모습으로 시각화된 모든 페이지들은 서로의 끌어당기는 힘에 의해 유기적으로 재배치되는 모습의 인터랙션을 제공한다. 서로 다른 사이트는 구성 페이지들의 연관관계에 따라서 서로 다른 태양계의 모습으로 표현될 것이다. 결국 이 연구는 사용자에게 웹사이트의 대략적인 성격을 파악하는 것에 도움을 주고 웹 사이트 안에서의 페이지 탐색 시, 관련 주제의 정보가 속해있는 비슷한 페이지 들을 알려 줌으로써 보다 효율적인 정보 검색을 돕는다.

핵심어: sitemap, visualization, hyperlink, node-link diagram, relational analysis, gravity

1. 서론

1.1 연구 배경

현재 인터넷은 거듭되는 발전으로 인하여 엄청나게 많은 사이트들이 생겨났다. 이 사이트들은 타 사이트와의 경쟁에서 이기고 보다 많은 사용자를 끌어들이기 위하여 더 많은 양의 정보를 유저에게 전해 주려는 노력하고 있다. 이로 인해 생기는 웹 페이지 수의 증가는 웹 페이지 서로의 hyperlink 연결이 그물처럼 복잡해 지게 되는 원인이 된다. 물론 많은 페이지 수로 인하여 사용자는 많은 정보를 제공받을 수 있지만, 이는 또한 자신에게 유용한 정보를 찾기 어려운 문제점을 야기 시킨다.

1.2 연구 목적

이 연구에서는 특정 사이트 내에서 제공하고 있는 정보를 사용자가 직관적으로 파악할 수 있게 시각화 하고 그로 인해 시각적인 즐거움 까지 주는 것을 목적으로 하고 있다.

1.3 논문 구성

이 논문은 연구의 목적과 문제들의 해결 과정들을 정리하여 총 6 개의 단락으로 구성되며 다음과 같은 내용을 중심으로 진행된다.

1 장에서는 본 연구의 배경 및 필요성, 연구의 진행 개요에 대해 기술한다.

2 장에서는 본 연구의 발상과 구현에 배경이 되는 이론들을 살펴보게 된다.

3 장에서는 본 연구의 결과 제안하는 인터페이스와 인터랙션을 기술 하였다.

4 장에서는 시각화를 위한 정보의 수집과 분석 방법에 대하여 기술한다.

5 장에서는 지금까지의 연구를 구현하는 프로토타입 제작 과정을 기술하였다.

마지막으로 6 장에서는 본 연구의 방법과 내용, 결과를 종합하고 이를 바탕으로 향후 연구 과제와 발전 가능성을 모색한다.

2. 연구의 배경 이론

2.1 직관적인 시각화 이론

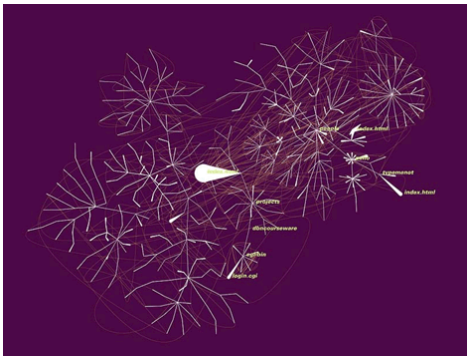


그림 1 Courtesy of Ben Fry, Aesthetics and Computation Group, Media Lab, MIT

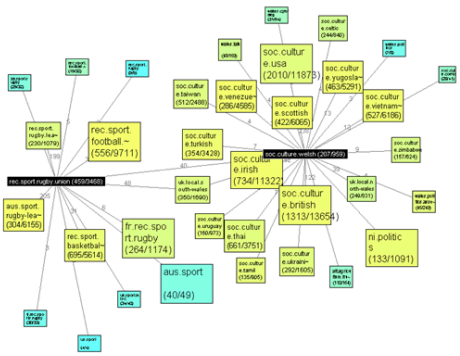


그림 2 Courtesy of Marc Smith, Microsoft Research

정보의 양이 많아 지는 것에 대하여 이 정보들을 직관적으로 표현 해보려는 노력은 많은 연구자들에 의해서 계속해서 진행되고 있다. 이러한 연구 중의 하나로 정보를 지도화해서 표현하는 것이다.

그림 1 은 웹 사이트의 사이트 맵을 사용자의 이용률을 토대로 시각화를 하였다. 이 때 각 페이지의 연결 모습을 노드 다이어그램으로 표현하였고 이러한 모습은 아네모네 꽃의 모습에 비유하였다. 사용자의 페이지 이용 로그를 분석하여 이용률에 따라서 선의 두께를 다르게 표시해서 어떤페이지가 이용률이 많은지를 직관적으로 알 수 있게 하였다.

그림 2 는 구독 중인 news 에 올라오는 게시물들의 연관관계를 분석하여 시각화 하였다. 이러한 시각화는 비슷한 성격의 주제들을 한눈에 알아 볼 수 있게 도와 준다.

이러한 노드 링크 다이어그램에 의한 시각화는 표현하고자 하는 정보를 사용자가 직관적으로 확인 할 수 있도록 하는 데에 도움을 준다. 이때 대부분의 사용자들은 각 노드가 링크에 의해 연결되어 있거나 서로 근접해 있을수록 관련성이 높다고 판단하게 되며 노드의 색이 진하거나 크기가 클수록 더 큰 영향력을 행사하고 있는 노드라고 판단한다.

2.2 Screen Scrapping

2.2.1 Screen Scrapping 의 개요

Screen Scrapping 은 화면에 표시되는 웹 페이지를 구성하는 html 태그 정보를 분석해서 원하는 정보를 뽑아내는 기술이다. 서로 다른 페이지에서 추출한 데이터를 이용하여 데이터 베이스로 활용하기 때문에 비교 분석 자료로 활용이 가능하고 추출된 데이터를 가공하여 새로운 형태로 보여 줄 수 있다.

이 기술을 사용하면 일일이 각 웹사이트를 방문할 필요가 없이 자료를 얻을 수 있기 때문에 시간과 경비를 절약할 수 있는 장점이 있으나 사생활 침해가 우려되며, 사이트가 갱신될 때마다 프로그래밍을 수정해야 한다는 단점도 가지고 있다.

2.2.2 Screen Scrapping 의 활용 사례

미국 등 선진국에서는 1990 년대 말부터 보편화 되었으며, 우리나라에는 2000 년 12 월 설립된 핑거(finger)사가 이 기술을 처음 개발하여 금융 서비스를 시작한 후 금융기관을 중심으로 널리 사용되고 있다. 현재 대표적인 것으로는 개인이 가진 여러 금융기관의 계좌들을 통합하여 관리할 수 있는 금융 자산 통합 관리 소프트웨어(PFMS: Personal Finance Management Software)를 들 수 있다.

PFMS 에서는 개인이 등록한 각 금융 사이트들을 프로그램에서 자동으로 로그인 하여 계좌의 잔고와 출금 내역에 대한 데이터를 읽어 오고 이 데이터를 토대로 개인 총 자산이 얼마인지 출금은 얼만큼 했는지를 파악할 수 있도록 돕는다. Screen Scrapping 기술 없이 이런 기능을 실현하기 위해서는 각 금융 사이트들의 업무 협약과 통합 작업이 있어야만 가능할 것이었다.

2.3 관련 물리 이론

보다 직관적이고 자연스러운 시각화를 위해서 자연 법칙에서 사용되는 공식들을 인용하였다.

2.3.1 만유 인력의 법칙 공식

$F=Gmm'/r^2$

(F=인력, G=만유인력상수, m=한 물체의 단위 질량, m'=다른 한 물체의 단위질량, r=두물체사이의 단위거리)

서로 다른 물체 사이의 끌어 당기는 힘을 구하는 공식이다. 이 공식은 각 페이지의 연관성을 결정하는 공식으로 이용된다.

2.3.2 삼각함수

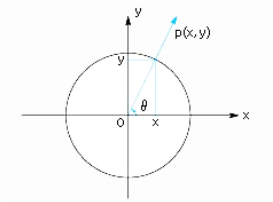


그림 3 삼각함수

평면에 O 를 원점으로 하는 좌표계를 정하고 이 평면 위의 점의 좌표를 (x, y)로 표시하고, x 축의 양의 방향에 대하여 각 θ 를 만드는 사선 OP 를 그어 O 를 중심으로 하는 단위 원과의 교점을 P 로 하여, P 의 좌표를 (x, y)라면, θ 가 주어질 때마다 x,y 가 정해진다.

이 공식은 연관된 페이지의 시각화 시 위치 결정에 사용된다.

3. 제안하는 인터페이스와 인터랙션

3.1 인터페이스

그림 4 태양계의 메타포

시각화의 방법으로써 태양계의 메타포를 도입하였다. 태양계의 각 행성들은 서로가 끌어 당기는 중력에 의해 위성이 되거나 다른 행성을 위성으로 삼는다. 또한 이러한 끌어 당기는 힘은 유기적으로 모든 행성에 영향을 미치는데, 이러한 상호 관계를 통해 만들어 지는 행성의 크기나 위치가 결정되는 모습을 웹사이트 페이지의 크기 표현과 위치 결정에 이용하였다.

3.2 인터랙션

웹사이트를 구성하고 있는 하나의 페이지는 하나의 행성으로써 서로 다른 크기의 원으로 표현한다. 각 페이지의 하이퍼링크를 통해 연결된 링크의 수를 통하여 행성 사이의 끌어 당기는 힘을 결정한다. 결국은 이런 힘으로 행성이 위성을 갖듯이 관련 성이 높은 페이지들끼리 모이게 되어 하나의 태양계의 모습을 이루게 되고, 페이지가 바뀌게 되면 서로의 당기는 힘이 바뀌게 되어 유기적으로 재배치 된다.

4. 시각화를 위한 정보의 수집과 분석

4.1 페이지 정보의 수집

페이지의 정보 수집과 처리 과정은 다음과 같으며, Screen Scrapping 시에 사용되는 처리 과정을 이용했음을 밝힌다. 우선 목표가 되는 웹사이트의 첫 페이지로부터 HTML 소스 분석을 한다. 소스상에 다른 페이지로의 링크 정보는 〈a href="~"〉의 형태로 표현되어 있다. 이 정보로부터 다른 페이지의 주소를 알아낼 수 있으며 다른 페이지에서 다시 소스 분석을 하여 다른 페이지들의 주소를 알아내는 작업을 반복하게 되며 각 페이지 정보로부터 알아낸 정보는 Database 화 되어 기록한다.

4.2 페이지 정보의 DB 화

4.2.1 Database 설계

테이블 - dbo.PAGE 요약

열 이름	데이터 형식	Null 허용
id	int	☐
url_addr	varchar(255)	☐
title	varchar(255)	☐
isvalidated	int	☐
		☐

그림 5 PAGE 테이블

테이블 - dbo.PAGERELATION 요약

열 이름	데이터 형식	Null 허용
parent_page_id	int	☐
child_page_id	int	☐
		☐

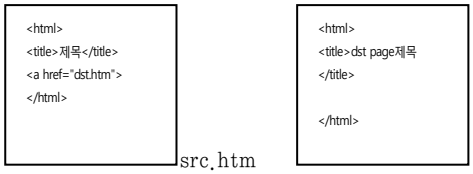
그림 6 PAGE 들의 관계를 저장 한 테이블

두 개의 테이블로 설계가 되었다. PAGE(id, url_addr, title, isvalidated)와 PAGERELATION(parent_page_id, child_page_id)가 그것이다.

PAGE 테이블은 저장 된 각 record 당 하나의 웹 페이지 정보를 기록한다. id 에는 각 웹 페이지의 고유 식별 ID 가 기록된다. url_addr 에는 해댕 웹 페이지에 접속하기위한 URL 이 기록된다. title 에는 페이지의 제목이 기록된다. isvalidated 컬럼은 해당 웹 페이지를 분석 시 오류가 있었는지의 여부를 기록하였다.

PAGERELATION 테이블은 PAGE 테이블에 축적되어 있는 웹 페이지 정보들의 관계를 명시하였다. PAGERELATION(parent_page_id, child_page_id)에서 parent_page_id 에는 부모 페이지의 id, child_page_id 에는 보모 페이지에 포함되어 있는 자식 페이지의 id 가 기록된다. parent_page_id 와 child_page_id 값에 의해서 각 페이지가 어떤 페이지에서 링크가 걸려 있는지를 몇 개의 링크를 가지고 있는지 등의 정보를 알아 낼 수가 있다.

4.2.2 Database 의 구축



src.htm

dst.htm

웹 페이지의 HTML 소스 정보를 분석하여 Database 를 구축하게 된다. 위 그림의 두페이지는 src.htm 에서 분석을 시작하여 dst.htm 페이지까지 연결되어 분석하게 된다.

Database 구축은 PAGE 테이블의 url_addr 에 src.htm 이 들어가게 되고 title 에는 〈title〉태그로부터 추출한 "제목" 그리고 src.htm 의 분석시에 오류가 없었다면 isvalidated 에 1(오류가 있는경우 0)가 들어간다.

이어서 〈a〉태그에 의해서 연결된 dst.htm 페이지를 같은 방법으로 분석하게 되며, 또다시 url_addr 에는 dst.htm, title 컬럼에는 "dst page 제목" 그리고 dst.htm 의 분석에서 오류가 없을 경우 isvalidated 값이 1 이 들어간다.

PAGERELATION 테이블의 parent_page_id 에는 src.htm 의 고유 번호가 할당 되고 child_page_id 에는 dst.htm 의 고유 번호가 할당 되어 새로운 record 가 추가 되게 된다.

4.3 페이지 정보의 활용

Database 에 구축된 내용은 시각화의 근거 자료로써 활용이 된다. PAGE 테이블의 정보는 시각화 시에 각 웹 페이지의 이름과 연결 주소를 알려 주는데 이용을 하며, 각 페이지와의 관계는 PAGERELATION 테이블의 정보로 부터 분석해 낸다. 이때 사용되는 SQL 쿼리는 다음과 같다.

(1) SELECT count(*) FROM PAGE WHERE parent_page_id=[해당페이지의 고유번호]

(2) SELECT count(*) FROM PAGERELATION WHERE child_page_id=[해당 페이지에 포함된 웹 페이지의 고유번호]

(3) SELECT * FROM PAGERELATION WHERE parent_page_id=[해당페이지의 고유번호]

(4) SELECT * FROM PAGERELATION WHERE child_page_id=[해당 페이지에 포함된 웹 페이지의 고유번호]

(1)의 쿼리로 자신의 자식 페이지 수를 구해서 중량을 구할 수 있고 (2)의 쿼리로 자식 페이지와의 관계 비중을 구할 수 있다. 또한 (3)의 쿼리로 같은 부모 페이지를 가지고 있는 모든 자식 페이지를 알아내고, (4)의 쿼리로 같은 자식들에 연관되어 있는 모든 페이지들을 알아낼 수 있다.

5. 프로토타입 제작

5.1 서버 환경 구성

본 연구의 구현에서는 유저의 요청을 받아 웹 페이지 정보를 수집하고 분석하는 역할을 하는 서버와, 서버에서 수집된 데이터를 보여 주는 역할을 하는 클라이언트로 나뉘어 진다.

프로토타입 제작에 사용된 서버의 구성은 다음과 같으며, 웹 페이지 수집 로봇을 설치와 Database 의 xml 변환 asp 프로그램을 설치하여 서버 구성을 완성하게 된다.

서버에 설치되야 할 프로그램은 MS Windows 2003 Server, IIS6.0, .net framework 2.0, MS-SQL 2003 Server 이다.

5.2 웹 페이지 수집 로봇

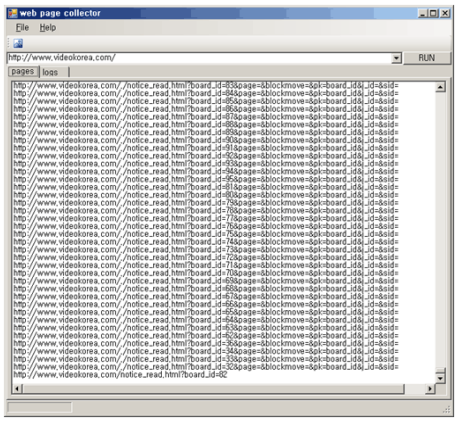


그림 7 수집 로봇

서버 측에서는 클라이언트로부터 요청을 받은 사이트의 페이지 정보를 추출 및 분석하여 Database 화 하기 위한 소프트웨어가 필요하다. 여기서는 이러한 역할을 하는 소프트웨어를 수집 로봇이라고 명했고 다음과 같이 역할을 수행할 수 있어야 한다.

첫째로 웹 페이지의 html 소스를 분석할 수 있어야 한다. 특히 이때 분석한 내용 중 〈title〉 태그로부터 페이지 제목을 추출할 수 있어야 하고, 〈a href="URL"〉 등의 태그로부터 연관된 페이지 주소를 추출 할 수 있어야 한다.

페이지의 주소는 절대 경로이던 상대 경로이던 간에 무조건 절대 경로로 변환하여 DB 화한다.

둘째로 수집 로봇은 두 개 이상의 사이트를 대상으로 동시에 정보 수집을 수행하여야 한다. 이를 위해서는 기존에 축적해 놓은 데이터를 재사용 할 수 있는 것이 효율 향상에 도움이 되고 수많은 요청에 대응하는 스레드를 관리하기 위한 매니저 프로그램의 개발이 필수이다.

동시 작업이 불가능 하다면 정보를 요구하는 사용자들은 이전 사이트의 수집 작업이 끝날 때까지 지루하게 기다려야만 한다.

셋째로 수집 로봇은 FLASH 등의 클라이언트 측에서 시각화를 담당하는 프로그램의 메시지를 전달 받을 수 있어야만 한다. 최소한 사용자가 시각화 하기를 원하는 페이지의 주소는 알아야 수집 작업을 시작하기 때문이다.

5.3 시각화 구현

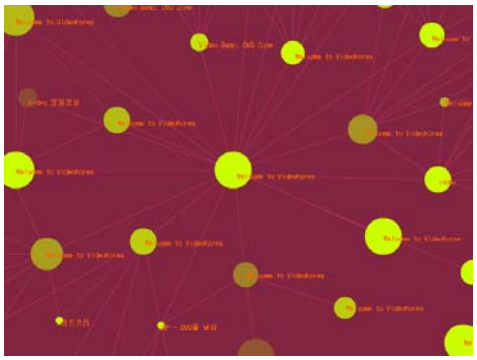


그림 8 시각화 구현 예제

본 논문의 시각화 프로토타입 구현에 Adobe Flash CS3 를 사용하였다. 이때 서버의 Database 와 연동을 위하여 Database 내용을 XML 로 전달 받기 위한 별도의 ASP 프로그램이 개발되었다.

그림 8 은 한 영화 정보 사이트를 대상으로 수집 로봇이 추출한 데이터를 토대로 시각화 된 모습이다. 구현에서 보듯이 노드-링크 다이어그램의 형태로 표현이 되었다.

각 노드는 하나의 웹 페이지를 나타낸다. 이때 가장 밝은 노드는 다른 페이지에서 연결된 수가 많음을 뜻한다. 그와 반대로 흐리게 표시되는 노드는 다른 페이지로부터의 연결 수가 적음을 뜻한다. 이는 곳 각 페이지의 이용률로 해석된다.

각 노드는 연관관계에 따라 링크로써 연결을 표현하였다. 연결되는 자식 노드가 많을 수록 노드의 크기는 커지게 된다. (여기서는 max 지름 50, min 지름 10 으로 세팅 되었다.) 다른 노드를 적게 가지고 있는 부모 노드의 경우는 작은 크기의 원으로 표현된다.

각 노드의 관계를 파악하는 것에 중요한 것은 각 노드의 인접도이다. 근처에 위치한 노드는 서로 연결이나 자식들의 연결이 비교적 빈번하게 일어났다는 것을 뜻하며 이는 곧 관련성이 높은 페이지일 가능 성이 높다는 것을 뜻한다. 예제 구현 결과는 거의 모든 페이지들이 사이트의 첫 페이지인 index.htm 에 연관되기 때문에 index.htm 이 모든 노드의 중심에 위치하게 되었다. 또한 신작 영화 소개의 관련 페이지가 뉴스 정보 페이지와 인접해서 관련성이 높다는 것을 확인 할 수 있었다.

6. 결론

6.1 결론

웹사이트의 복잡하게 이어져 있는 정보의 효율적인 제공에 대한 연구는 계속해서 이루어 지고 있으며 발전되고 있다. 이러한 연구는 본 논문에서 제시한 시각화 뿐 아니라 검색 방법이나 새로운 인터페이스 또는 효율적인 메뉴 구성에 의해서도 시도되고 있다.

본 연구는 이러한 연구 중의 하나로 사용자가 직관적으로 웹사이트의 모습을 파악하고 관련 정보가 있는 페이지로 연결할 수 있는 시각화를 시도하였다. 이를 위해서 대부분의 사용자들이 수긍하고 쉽게 받아들일 수 있는 규칙들을 조사해서 만들어 내었고, 이를 구현하기 위한 여러 관련 기술들과 메타포를 조사하였다.

이 연구에서는 또한 사회 인문학적인 조사, 분석과 더불어 최신 기술의 충분한 활용 그리고 직관적인 디자인과 새로운 볼거리를 제공하는 인터랙션까지 많은 부분의 연구 결과를 통합적으로 활용하려고 노력하였다.

그 결과 웹사이트를 구성하는 페이지들의 링크 관련성에 의해 만들어 진 노드-링크 다이어그램은 사용자에게 웹사이트의 규모와 구조를 파악하는 것과 더불어 관련 정보가 있는 페이지를 더 쉽게 찾게 도와는 역할을 하게되었다. 또한 해당 사이트가 편중되어서 다루고 있는 주제나 분야에 대한 대략적인 파악을 할 수 있으므로 운영자는 사이트의 취약한 부분을 분석할 수 있다.

이러한 시각화는 기존의 웹사이트의 전체 구조를 파악하기 위해 이용되던 사이트맵의 역할을 대신할 수 있을 것이라고 예상된다.

6.2 향후 연구 과제

본 연구에서 얻은 발견점과 한계점을 바탕으로 다음의 향후 연구 과제를 제시한다.

6.2.1 페이지 정보 추출 속도 개선

본 연구에서 제시한 Screen Scrapping 기법은 HTML 웹 페이지의 소스를 분석하기 때문에 HTML 페이지 전체를 다운 받는 작업이 선행이 된다. 이 때문에 페이지마다 접근하는데 적지 않은 시간이 걸린다.

테스트를 통해 대규모의 사이트일 경우 소속 페이지들을 수집하고 분석하는 대에 굉장히 오랜 시간이 걸리는 것을 확인 하였다.

현재는 이러한 문제를 분석이 된 정보를 실시간으로 보여 줌으로 써 어느 정도 해결하였지만 그래도 만족하지 못하는 속도는 해결해야 할 과제 중 하나이다.

6.2.2 웹 페이지 이외의 연결에 대한 지원

현재는 HTML 소스 분석에 의해서만 연결된 페이지를 수집하고 있다. 하지만 최근의 웹 경향상 FLASH 나 ActiveX 등으로 구현되어 있는 페이지가 늘고 있다는 것은 부인 할 수 없는 사실이다.

지금은 이러한 EMBEDDED 프로그램에서 연결된 웹 페이지는 알아내지 못하고 있지만 보다 정확한 시각화를 위해서는 모든 링크 정보를 알아내야만 할 것이다.

6.2.3 시각화 결과의 User Interface 강화

시각화 시에 User Interface 는 간과할 수 없는 문제이다. 특히 본 연구에서는 매우 많은 노드를 제한된 화면에 표현해야 하기 때문에 보다 효율적인 User Interface 가 연구되어야 한다.

6.2.4 시각화 Interaction 의 강화

본 연구의 목적은 시각화에 의한 정보 전달 이외에도 유저에게 새로운 즐거움을 주는 것을 가지고 있었다. 또한 사용자의 조작에 의한 상호작용 시에 주요 정보를 알려주는 인터랙션이 필요하다고 생각하였다. 이를 위한 인터랙션에 대한 연구가 필요하다.

6.2.5 보다 직관적인 디자인

향후에는 보다 직관적인 시각화를 위하여 과학적이고 근거가 있는 노드 색의 선택과 노드 모양을 이용해야 할 것이다. 또한 이때 추출된 시각화 모습은 웹사이트의 개관적인 모습과 특징적인 성격을 표현할 수 있는 지표로 까지 발전될 것이다.

참고문헌

[1] Ben Fry, Aesthetics and Computation Group, Media Lab, MIT, “Anemone visualization”.

[2] Marc Smith, Microsoft Research, “Mapping the Social Geography of Usenet News”.

[3] 이명로, 정지홍, “웹 사용 데이터를 이용한 사이트맵 시각화에 관한 연구”.