컴포넌트 기반의 웹 로그 분석 시스템 설계

심민석[0] 유대승 엄정섭 이명재

울산대학교 컴퓨터 · 정보통신공학부

{wolf, yds, quaza, ymj}@cic.ulsan.ac.kr

Design of Component-Base Analysis System

Minsuck Sim[O] Daesung Yoo Jeongseob Eom Myeongjae Yi

School of Computer Engineering & Information Technology, University of Ulsan

요 약

웹의 빠른 발전과 더불어 다양한 시스템들이 웹 기반으로 통합되고 있으며 복잡해진 시스템을 개발 및 유지보수하기 위해서 웹 분석 도구들의 필요성이 증대되고 있다. 웹 분석 도구들 중 웹 로그 분석 도구들은 축적된 웹 로그를 분석하여 분석 가능한 정보를 산출하고 이를 효율적인 웹 운영의 기초 정보로 이용할 수 있도록 한다. 그러나 기존의 웹 로그 분석 도구들은 웹사이트에 따라 요구되는 로그 분석 목적을 충족시키기 위해서 변형되어야 하거나 특정한 목적에 맞게 새롭게 개발되어야 한다. 본 연구에서는 로그 분석 시스템 또는 로그 분석 도구 개발에서 공통적으로 수행되는 과정(로그 파일로부터 필요한 항목들을 추출하고 정제하여 분석 가능한 데이터로 저장 루틴)을 컴포넌트화하였다. 이 컴포넌트는 로그로부터 추출한 정보를 XML 문서 또는 데이터베이스에 저장한다. 본 연구에서 개발한 컴포넌트는 로그 데이터를 XML 문서 형식과 데이터베이스에 로그 정보를 저장하여 쉽게 다른 시스템에서 저장된 정보를 유용하게 사용하도록 한다. 또한 생성한 컴포넌트의 효용성을 검증하기 위하여 웹 로그 분석 시스템을 설계하였다.

1. 서 론

웹의 발전과 더불어 웹 기반 소프트웨어의 개발 및 유지보수 패러다임이 변하고 있다. 그 예로는 짧은 소프트웨어의 생명주기와 시스템의 복잡함을 들 수 있다. 이러한 변화로 인하여 시스템을 개발하고 유지보수하는 데 보다 많은 비용이 소요된다[1,2]. 이에 효율적으로 웹 기반의 시스템을 개발 및 유지보수하기 위해서는 웹 기반 시스템의 분석에 관한 연구가 필요하다.

일반적으로 로그 분석은 로그 정보를 분석하는 것과 로그 정보와 그 외 다양한 정보를 상호 연동하여 분석하는 좀더 확장된 것으로 분류한다. 일반적 의미의 로그 분석은 로그 데이터를 이용하여 트래픽을 파악하고, 이 트래픽이 지닌 의미를 분석해 나가는 것이라고 할 수 있다. 로그 데이터를 이용하여 웹사이트의 페이지뷰, 사용자별 페이지뷰, 접속장소 및 방식, 시간별 페이지뷰, 방문자수 등에 대한 현황 및 추세를 분석하는 것이다. 그래서 사용자가 웹사이트를 방문하는 경로와 서핑하는 경로에 대한 분석을 통하여 웹사이트가 지닌 문제점을 찾고, 사용자가 웹사이트에서 무엇을 원하는지를 보다 구체적으로 파악하는 것이다. 이에 반해, 확장된 의미의 로그 분석은 단지 로그 데이터뿐만 아니라, 웹사이트에서 보유하고 있는 고객 등록정보, 구매정보, 외부환경정보 등을 복합적으로 사용하는 것을 말한다. 이러한 분석을 통하여 사용자 특성별로 웹사이트의 이용, 구매에 대한 보다 폭넓은 분석이 가능하며 로그 분석을 통해 웹사이트는 사용자에 대해 보다 정확히 파악할 수 있게 된다. 그러나 로그 데이터는 그 자체로만 분석할 수 있는 정보가 그다지 많지 않으며 금융권의 데이터와는 달리 부정확하다는 특징을 가지고 있다.

인터넷과 관련된 기업의 폭발적인 증가에 비해 성공적인 웹사이트 운영에 필수적인 로그 분석의 수준은 초보적인 수준이라고 말할 수 있다. 웹 로그 분석을 단지 방문자수, 페이지뷰 정도의 정보로밖에 인식하고 있는 것이 일반적인 상황이고 심지어 웹 로그 분석에 대한 개념 조차도 가지고 있지 않는 곳이 있다.

웹 로그 분석의 가장 일반적인 방법은 웹 로그 분석 도구를 이용하는 것이다. 잘 알려진 웹 분석 도구는 상용화된 WebViz, WebTrends, CountBoy 등이 있고 웹 서버 자체적으로 제공하는 분석 도구가 있다. 분석 도구를 이용할 경우, 접속자수와 방문자수, 방문자들의 분류, 방문객의 접속 ISP별 집계, 홈페이지 디렉토리와 파일별 통계, 시간별(월 · 주 · 요일 · 일 · 시간) 분석 등의 정보를 알 수 있다. 하지만 이들 분석 도구는 웹사이트를 구성하고 있는 디렉토리와 파일 이름을 기반으로 처리하기 때문에 웹사이트의 디렉토리나 파일 이름들을 분석하기 위해 구조화하지 않을 경우에는 그 효용 가치가 떨어질 수밖에 없다. 그리고 디렉토리나 파일 이름이 무엇을 의미하는지 정확히 알고 있지 않을 경우에도 의미없는 데이터가 될 수도 있다. 또 다른 문제점은 이들 분석 도구는 분석에 불필요한 파일들을 정확히 필터링할 수 있는 기능들을 대부분 지원하지 않기 때문에 그런 불필요한 파일로 인해 정보의 정확성은 떨어질 수 있다. 이 경우 정확한 페이지뷰를 파악하기 위해서는 별도의 작업이 필요하게 된다. 그리고 이들 분석 도구를 이용할 경우 한 가지 관점에서만 분석할 수밖에 없다는 한계를 가지고 있다. 방문자들, 페이지 내용, 시간별 분석에 있어 서로 연관된 종합적인 분석을 할 수 없기 때문에 이 경우는 다른 형태의 로그 분석을 고려해 봐야 한다. 따라서 기존의 웹 로그 분석 도구를 이용한다 하더라도 웹사이트의 분석 목적에 맞는 분석 알고리즘을 내부적으로 가지고 있어야 한다. 그러므로 보다 정확하고 가치있는 정보를 얻기 위해서는 웹사이트에 맞는 자체적인 분석 도구를 개발하는 것이 좋은 방법일 수 있다. 하지만 자체 분석 도구를 개발하는 것은 쉬운 작업이 아니다.

본 논문은 웹 분석 과정의 공통적인 부분을 컴포넌트로 만들어서 웹 도구의 자체 개발을 쉽고 편리하게 구성하도록 만들었다. 웹 분석 과정을 크게 로그 데이터를 분석 가능한 데이터로 변경하는 부분과 분석 가능한 데이터를 웹사이트에 맞는 알고리즘을 사용하여 데이터를 가공하는 부분으로 분류하고 이에

본 연구는 정보통신부의 "정보통신 우수시범학교 지원사업"의 지원에 의해 이루어졌습니다.

해당하는 컴포넌트를 작성한다. 이렇게 작성한 컴포넌트의 효용성을 검정하기 위하여 간단한 웹 분석 도구(그림3)를 설계한다.

따라서 웹사이트 특성에 맞는 분석 도구를 개발하여 이용하는 것은 기존의 웹 로그 분석 도구를 사용할 때 보다 정확한 분석을 할 수 있고 로그 분석 도구를 구현하는 일반적인 방법에 비해서 빠르고 쉽게 설계 및 구현을 할 수 있다.

본 논문의 구성은 다음과 같다. 2장은 기존의 웹 분석 시스템과 웹 테스트 시스템에 대해서 알아본다. 3장은 웹 로그 분석 컴포넌트들에 대해 살펴보고 4장에서는 제안한 컴포넌트를 조합하여 간단한 웹 로그 분석 시스템을 설계한다. 5장에서는 결론 및 향후 연구 방향에 대하여 논한다.

2. 관련 연구

웹 로그 분석에 관련된 국내의 연구 동향으로는 아직까지 국외에 비해 활성화되어 있지 못한 상태이다. 국외에서는 웹 분석에 대한 연구가 활성화되어 웹 분석을 위한 많은 시스템들이 출시되고 있다. 잘 알려진 웹 분석 도구들 중에는 WebViz, WebTrends, CountBoy가 있다.

WebViz[3]은 Unix 환경의 프로그램으로 웹 로그를 분석하여 사용자, IP, 기간별, 웹페이지별 접근이 가능하며 웹페이지와 페이지간의 링크를 비주얼하게 보여주는 도구이다. 링크의 빈도수 표시는 RGB 칼러를 사용하여 표현하였다.

WebTrends[4]은 접속한 페이지, 사용자, 특정 웹 브라우즈, 인증 등 기타 사항을 그래프 형식으로 보여주는 도구이다.

CountBoy[5]은 로그 분석을 원하는 페이지에 간단한 스크립트를 삽입하여 필요한 데이터를 추출하고 분석하는 도구이다.

3. 웹 로그 분석 컴포넌트

본 논문은 웹 로그를 분석하는 컴포넌트를 크게 2가지 영역(로그를 분석하여 분석 가능한 데이터로 나누는 영역, 분석 가능한 데이터를 활용하는 영역)으로 분류한다. 분석 가능한 데이터를 생성하는 컴포넌트는 WebLog2XML 컴포넌트 등이 있고 분석 가능한 데이터를 활용하는 컴포넌트는 WebLogViz 컴포넌트 등이 있다.

WebLog2XML 컴포넌트는 다양한 웹 로그 데이터(NCSA, IIS, 확장 IIS) 형식에서 XML 기반의 분석 가능한 공용 데이터 형식의 데이터로 변경하는 기능을 가지고 있으며 아래와 같은 인터페이스를 가진다.

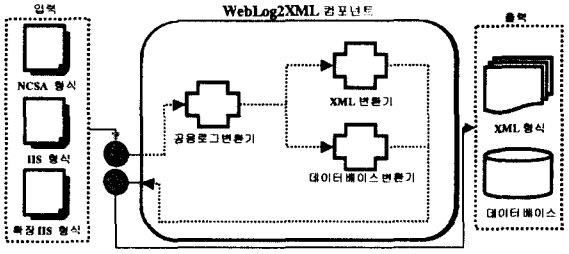

그림 1 WebLog2XML 컴포넌트 구조

WebLog2XML 컴포넌트는 IWebLog2XML 인터페이스와 컴포넌트의 내부 에러 처리를 위해 _IWebLog2XMLEvents 인터페이스를 제공한다.

표 1 WebLog2XML 인터페이스

```
interface IWebLog2XML : IDispatch
{
    HRESULT Load_LogFile(BSTR       m_Data);

    HRESULT Translate_NCSA();
    HRESULT GetLostData(BSTR        *m_LostData);

    HRESULT NCSA2XML(BSTR           *m_PureData);
    HRESULT NCSA2DB(BSTR            m_ODBC,
                    BSTR            m_name,
                    BSTR            m_passwd);
};

dispinterface _IWebLog2XMLEvents
{
    HRESULT _Err(int err_num, BSTR m_Desc);
};
```

IWebLog2XML의 Load_LogFile 메서드는 다양한 타입 형식의 로그 데이터를 읽어 파일의 종류 및 시간 등의 정보를 사용하여 분석 환경을 설정한다. Translate_NCSA 메서드를 사용하여 공용 로그 형식(그림 4)으로 변경하며 변경된 결과를 분석 프로그램에서 사용할 수 있도록 이용 가능한 데이터 형식으로 바꾸는 NCSA2XML, NCSA2DB 메서드를 제공한다. 또한 공용 로그 형식으로 변경하면서 손실된 데이터를 저장하기 위해 GetLostData 메서드를 제공한다.

WebLogViz 컴포넌트는 이용 가능한 데이터(XML, 데이터베이스) 형식에서 문서 구조 정보와 페이지 히트 정보를 비주얼하게 보여주는 기능을 가지고 있으며 아래와 같은 인터페이스를 가진다.

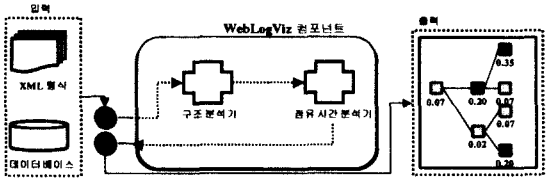

그림 2 WebLogViz 컴포넌트 구조

WebLogViz 컴포넌트는 IWebLogViz 인터페이스와 컴포넌트의 내부 에러 처리를 위해 _IWebLogVizEvents 인터페이스를 제공한다.

표 2 WebLogViz 인터페이스

```
interface IWebLogViz : IDispatch
{
    HRESULT Load_XMLFile(BSTR       m_Data);
    HRESULT Load_ODBCInfo(BSTR      m_ODBC,
                          BSTR      m_name,
                          BSTR      m_passwd);

    HRESULT Parser();
    HRESULT Cal_Occupation_Time();

    HRESULT WebLogViz();
};

dispinterface _IWebLogVizEvents
{
    HRESULT _Err(int err_num, BSTR m_Desc);
};
```

IWebLogViz의 Load_XMLFile, Load_ODBCInfo 메서드는

XML 형태 또는 데이터베이스에 저장된 공용 로그 형식 데이터를 읽어 문서 입력 정보를 설정한다. Parser 메서드는 공용 로그 필드 중 "Request Type"의 정보를 사용하여 문서 구조 정보를 생성하고 Cal_Occupation_ Time 메서드를 사용하여 문서 구조 정보 데이터에 사용자 점유 시간 정보를 기록한다. 마지막으로 WebLogViz 메서드는 추출한 구조 정보와 사용자의 페이지 점유 시간을 비주얼하게 표현한다.

4. 간단한 웹 로그 분석 시스템 설계

그림3는 WebLog2XML 컴포넌트를 이용하여 간단한 웹 로그 분석 도구 설계도이다. 다양한 로그를 사용하여 웹 컨텐트 구조를 추출하고 사용자 점유 시간을 표시하는 시스템이다.

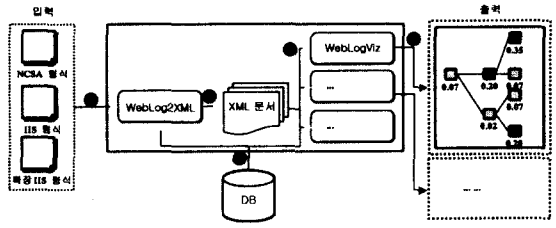

그림 3 웹 로그 분석 도구 구조

그림4는 WebLog2XML 컴포넌트를 사용하여 분석 가능한 공용 로그 형식이며 표3은 XML 문서 구조 정보이다.

그림 4 공용 로그 형식

표 3 공용 로그 저장을 위한 XML의 DTD

```
<!DOCTYPE WebLog[
   <!ELEMENT Logs          (item*)>
   <!ELEMENT item          (Remote-host-name,
                            Username,
                            Date,
                            Time-and-GMT-offset,
                            Request-type,
                            Service-Status-code,
                            Bytes-sent )>

   <!ELEMENT Remote-host-name     (#PCDATA)>
   <!ELEMENT username             (#PCDATA)>
   <!ELEMENT Date                 (#PCDATA)>
   <!ELEMENT Time-and-GMT-offset  (#PCDATA)>
   <!ELEMENT Request-type         (#PCDATA)>
   <!ELEMENT Service-Status-code  (#PCDATA)>
   <!ELEMENT Bytes-sent           (#PCDATA)>
]>
```

5. 결론 및 향후 연구 과제

본 논문은 웹 어플리케이션에 맞는 자체적인 분석 시스템을 개발하는 데 필요한 컴포넌트(WebLog2XML, WebLogViz)를 작성하고 조합하여 웹 분석 도구를 설계해 보았다.

본 논문에서 제안한 컴포넌트를 사용하면 특정 웹 어플리케이션에 맞는 웹 분석 어플리케이션의 작성이 용이하고, 웹 로그 분석 시스템의 종합적인 분석의 한계를 어느 정도 해결하였다.

향후 연구 과제로는 활용 가능한 데이터를 사용하여 사용자 패턴 추출 컴포넌트 외에 데이터를 가공하는 부분, 가공된 데이터를 차트로 보여주는 부분, 상용화된 마이닝 도구 입력 형식으로 변경하는 부분 등 다양한 종류의 컴포넌트를 작성하는 연구와 로그 분석에 따른 결과를 사용하여 쉽게 웹 어플리케이션을 유지보수에 관한 연구도 필요하다.

[참고문헌]

[1] Benoit Leger, Jean-Christophe Cimetiere, "Web Load and Performance Testing Tools", "www.trendmarkers.com", 2000

[2] Hung Q.Nguyen, "Testing Applications on the Web", Wiley Computer Publishing, 2001

[3] James E. Pitkow & Krishna A. Bharat, "WebViz: A Tool for WWW Access log Analysis"

[4] WebTrends, http://www.webtrends.co.kr/

[5] http://www.countboy.com/

[6] Parunak, H.Van Dyke, (1989) "Hypermedia topologies and user navigation", Hypertext Proceedings, 43-50, 1989

[7] Rivlin, Ehud & Botafogo, Rodrigi & Shneiderman, Ben,(1994). Navigating in hyperspace: designing a structure-based toolbox. Communications of the ACM 37, 2. 87-96

[8] J.A. Whittaker and M.G Thomason. A markov chain model for statistical software testing. IEEE Trans. on Software Engineering, 20(10): 812-824, Oct. 1994.

[9] J. Tian and A. Nguyen. statistical web testing and reliability analysis. in Proc. 9th Int. Conf. on Software Quality, pages 263-274, Cambridge, MA, Oct. 1999

[10] E. Nelson. Estimating software reliability form test data. Microelectronics and Reliability, 17(1) 67-73, Mar. 1993

[11] 서연규, 김경중, 정윤경, 조성배 "웹사이트의 구조분석을 위한 소프트웨어 에이전트" 정보과학회 학술발표 논문집 27권 2호, pp.21-23, 2000

[12] 이기열, 이병정, 이숙희, 우치수 "웹 애플리케이션을 위한 복잡도 척도" 정보과학회 학술발표 논문집 27권 2-1호, pp.421-423, 2000