FastMap을 이용한 웹 문서 시각화 시스템의 설계 및 구현

문진석*, 손기락, 김차성
한국외국어대학교 컴퓨터공학과

Design and Implementation of Web Document Visualization System using FastMap

Jinsuk Mun, Kirack Sohn, Chasung Kim
Dept. of Computer Science & Engineering, Hankuk University of Foreign Studies

요 약

인터넷의 발달과 더불어 매일같이 제공되는 수많은 정보로부터 자신에게 필요한 정보만을 추출하는 데는 많은 시간과 노력이 소요된다. 이러한 정보수집의 어려움에서 정보를 쉽고 효율적으로 찾기 위해서 웹 문서 시각화 시스템을 구현하였다. 웹 문서 시각화 시스템은 사용자가 검색하는 정보는 과거에 검색했던 웹 문서를 다시 방문하는 경험에서 착안하였다. 이를 위해 인터넷 익스플로러을 통해서 방문 중인 웹 문서의 URL, 키워드, 문서간의 유사성을 추출하여 시각화 한다. 시각화 알고리즘으로 FastMap을 사용하였다. 본 논문에서 FastMap은 웹 문서간의 유사성, 즉 상대적인 거리 객체 형태를 2-차원 공간으로 표현하는 알고리즘이다. 2차원 공간으로 매핑된 주변에 있는 웹 문서 객체들을 확대하면 방문중인 웹 문서와 유사성이 있는 문서를 쉽게 찾을 수 있다.

1. 서론

흔히들, 인터넷을 정보의 바다라고 말하며 많은 양의 정보들이 산재해 있다. 그러나, 방대한 정보의 양이 곧 많은 정보를 습득하고 이용할 수 있음을 보장하지 못하고 있다. 오히려, 정보량의 증가에 따른 정보과잉현상은 자신이 필요한 정보를 추출하는데 많은 시간과 노력을 요구한다. 이러한 환경에서 사용자가 방문한 웹 문서는 과거에 방문한 웹 문서를 다시 방문하는 경험에서 웹 히스토리 정보를 보다 체계적, 시각적으로 분류하는 도구가 필요하다[1].

본 논문에서 제시하는 웹 문서 시각화 시스템은 사용자가 방문한 웹 히스토리를 보다 체계적으로 관리하기 위해 웹을 방문할 때 마다 문서의 유사성을 추출하여 유사성이 있는 웹 문서끼리 시각화(Visualization) 하였다. 시각화를 위한 알고리즘으로 FastMap를 사용하여 웹 문서간의 유사성 거리를 2-차원 공간으로 매핑한다. FastMap는 웹 검색이나 클러스터링에서 문서간 유사성을 k-차원($k=1,2,3$)으로 매핑하는 문제나 n-차원($n>k$, $n>3$)을 k-차원으로 감소시키는 문제에서 적용되고 있다.

본 논문의 구성은 2절에서는 관련 연구로 벡터 공간 모델, MDS(Multi-Dimensional Scaling)를 기술하고, 3절에서 시스템 구성, 웹 문서 추출, 웹 문서 분석, 데이터베이스 설계, 시각화, 사용자 인터페이스를 다루였으며, 마지막 4절에서는 결론 및 향후 연구방향을 다룬다.

2. 관련 연구

2.1 벡터 공간 모델

정보검색 분야에서 문서간의 유사한 정도를 기술하기 위해 여러 모델을 개발하였다. 그 중에서 가장 대표적인 모델인 벡터 공간 모델은 문서간의 벡터를 n차원 벡터공간으로 생각하여 벡터 사이의 cosine 값을 측정하여 문서간에 유사한 정도를 계산하는 방법이다. 이 모델은 문서간의 유사성 검색 실험에서 대부분 기본모델로 사용되었으며, 특히 SMART 시스템실험에서 채택한 모델이다. 이 시스템은 연관피드백기법, 클러스터링 기법, 어휘분석 기법과 같은 정보검색의 여러 분야에서 실험 되였다. 이 실험을 통해서 문서내 용어빈도수와 코사인 측정법에 의해 정규화된 역문서빈도수(Inverted document frequency)를 조합하면 가장 좋은 문서 용어 가중치를 만들어 낼 수 있다는 것과 용어 가중치를 갖는 질의에 대해서 개선된 질의 용어 가중치 부여 방법을 사용하면 성능을 향상 시킬 수 있다는 것을 보여주었다[2].

2.2 MDS

MDS(Multi-Dimensional Scaling)는 FastMap과 같은 문제, 즉 유사성을 k-차원으로 매핑하는 문제나 n-차원을 k-차원으로 감소시키는 문제와 관련된 다양한 분야(eg.,사회 과학, 심리학, 시장 분석, 물리학)에서 적용되고 있다. 또한, MDS는 데이터들간의 유사성으로부터 데이터 집합의 구조를 밝히는데 유용하다.

반면에 어플리케이션 이용에 FastMap과 비교하여 두 가지 단점이 있다. 첫째, 데이터 N개의 수행 복잡도가 $O(N^2)$ 이다. 이는 큰 데이터 집합 즉, 대용량 데이터베이스에 적용에 비실용적이다. 둘째, 빠른 데이터 검색이 불가능하다. 데이터 베이스에서 항목의 추가, 검색등에 가장 적합한 수행 복잡도는 $O(N)$이다 그러나 MDS는 $O(N^2)$의 복잡도로 데이터를 효율적으로 검색 할 수 없다[3].

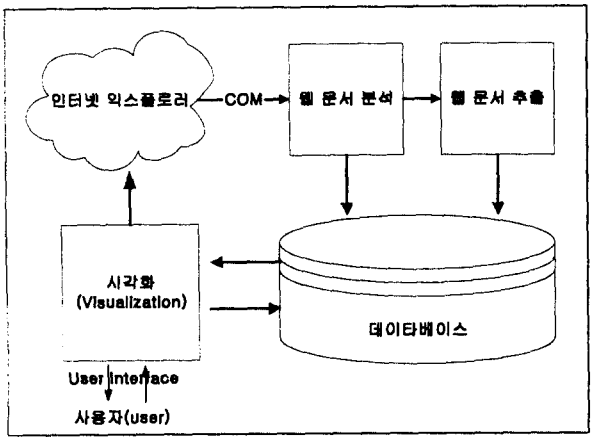
본 논문에서 이용한 FastMap은 수행 복잡도가 $O(Nk)$이므로

빠른 데이터베이스 검색이 가능하다.

3. 웹 문서 시각화 시스템

본 논문의 웹 문서 시각화 시스템의 전체적인 시스템 구성은 크게 두 부분으로 구성 되어 있다. 첫째, 웹 클라이언트인 인터넷 익스플로러에 웹 문서 시각화 시스템을 붙여 웹 문서 추출, 웹 문서 분석 그리고 데이터베이스에 저장하는 전처리 부분이다. 둘째, 전처리 부분에서 얻어진 정보를 이용하여 시각화, 사용자 인터페이스 부분이다.

웹 문서 시각화 시스템의 전체적인 구조는 <그림 1>과 같다.



<그림 1> 웹 문서 시각화 시스템 구조

3.1 웹 문서 추출

사용자가 웹 문서 시각화 시스템을 쉽게 사용할 수 있도록 하기 위해서 COM 을 사용하여 인터넷 익스플로러 4.0 에 붙였다. 웹 문서 추출은 인터넷 익스플로러 4.0 에서 현재 보고 있는 웹 문서의 다운로드가 완료되고, 파싱이 완료됨을 알리는 DocumentComplete 이벤트가 발생하면 웹 문서 추출모듈을 호출한다. 웹 문서 인터페이스를 얻기 위해서 MSHTML.DLL 의 IHTMLDocument2 인터페이스를 사용한다. 이 인터페이스는 일종의 트리 구조 컬렉션 객체이다. 즉 파싱된 HTML 의 객체들을 구조적으로 관리하는 역할을 한다. 따라서 IHTMLDocument2 인터페이스를 얻으면 파싱된 HTML 에 자유롭게 접근할 수 있다. IHTMLDocument2 는 듀얼 인터페이스로 구현된 거대한 자동화 객체이다. IHTMLDocument2 인터페이스를 얻는 방법은 첫째 자동화 메커니즘을 사용하는 방법과 둘째 직접 인터페이스 함수를 호출 하는 방법 두 가지 방법이 있다. 첫째 방법은 Visual Basic 과 같은 환경에서 쓰이며, 수행 속도면에서 직접 인터페이스 함수를 호출하는 방법에 비해 현저하게 떨어진다. 둘째 방법은 Visual C++ 환경에서 수행속도나 편이성에서 첫번째 방법 보다 좋다. 본 논문에서는 수행 속도의 향상을 위해서 인터페이스 함수를 호출하는 방법을 사용한다.

또한 IHTMLDocument2 는 매우 많은 프로퍼티와 메쏘드를 가지고 있으며 이중에 get_all() 프로퍼티를 사용하여 HTML 의 모든 요소들의 컬렉션 객체를 추출한다[4].

3.2 웹 문서 분석

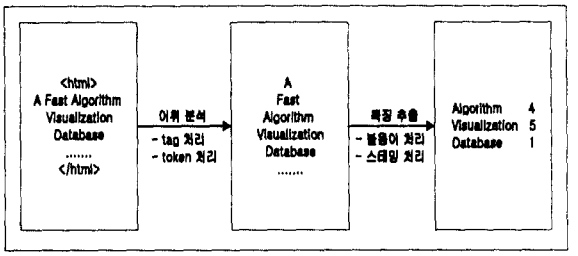
웹 문서 분석은 특징 추출(Feature Selection)과 문서간의 유사성을 추출하는 작업으로 구성하였다.

특징 추출은 임의의 문서에서 문서를 대표 할 수 있는 중요 키워드를 추출하여 문서의 특징이 되는 키워드 벡터를 형성하는 방법이다. 문서간의 유사성은 문서간의 키워드 벡터 사이의 코사인 값을 측정하여 문서간에 유사한 정도를 계산하는 방법이다

3.2.1 특징 추출

먼저 문서에 나오는 모든 문자들을 단어 또는 토큰 열로 변환하는 어휘 분석 과정을 수행한다. 그 결과 문서는 키워드와 출연빈도로 이루어진 벡터로 표현되고, 이렇게 표현된 벡터는 특징 추출을 거쳐, 문서의 특징이 되는 중요 키워드를 표현하는 벡터만을 추출한다. 그리고 키워드의 빈도에 따라 가중치 벡터를 적용한다.

본 시스템에서 사용하는 특징을 추출하는 과정은 크게 두가지 과정을 거친다. 첫째, 각 문서의 키워드 중에는 a, the, and 등 (430 여 단어)과 같은 발생 빈도가 높지만 문서의 특징으로서 가치가 없는 불용어를 삭제한다. 둘째, 스테밍을 이용한 방법으로 각 키워드의 어형론적인 변형을 찾는 방법을 제공하는 것이다. 위의 두가지 방법을 이용하면 키워드 파일의 크기를 줄여, 색인에 요구되는 메모리 용량을 줄일 수 있고, 검색 효율을 높일 수 있다[2][5][6].



<그림 2> 웹 문서 분석

3.2.2 웹 문서 유사성

본 논문에서 문서사이의 유사도 측정방법으로서 다음 <식 1>에 표현된 코사인값 측정법을 사용한다. 코사인 측정법은 비교 대상인 벡터가 희소행렬로 표현될 경우 계산량을 줄일 수 있는 장점이 있다.

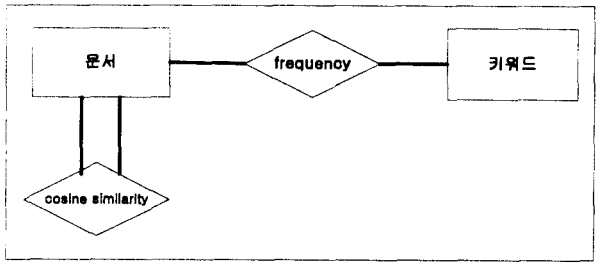
$$similarity(d1,d2)=\frac{\vec{u}1\circ\vec{u}2}{\|\vec{u}1\|\bullet\|\vec{u}2\|} \qquad (1)$$

'$\circ$'는 두 벡터의 내적이고, $\|*\|$ 는 벡터의 Euclidean norm 이다[3].

3.3 데이터베이스 설계

<그림 3>은 웹 문서 시각화 시스템의 시각화를 위해 설계한 데이터베이스의 E-R 다이어그램이다. 이는 시각화(Visualization)를 위한 기초 자료가 된다. 데이터베이스는 문서 테이블, 키워드 테이블, Cosine Similarity 테이블, Frequency 테이블로 구성 되며 문서 테이블은 문서번호, 문서의 웹사이트 주소인 URL, 문서의 키워드, 사용자에게 문서의 중요성을 나타내는 웹 문서의 방문 횟수, 2 차원 좌표 점으로 구성한다. 좌표 점은 문서간의

유사성을 FastMap 알고리즘에 적용하여 추출한 2 차원 좌표 점으로 새로운 문서를 방문할 때마다 구한다. 이를 데이터베이스에 저장하여 웹 문서 시각화 시스템의 수행 시간을 단축한다.



[그림 3] E-R 다이어그램

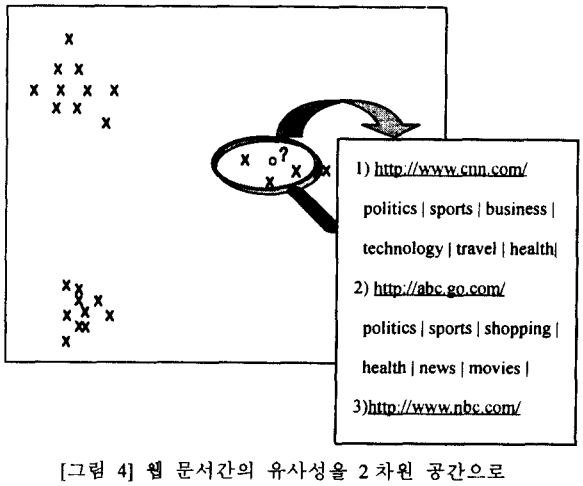
3.4 시각화

문서간의 유사성 거리 정보를 FastMap 알고리즘에 적용하여 시각화(Visualization)하는 모듈이다.

FastMap 알고리즘은 벡터 공간 모델에서 구한 유사성의 거리를 k-차원으로, n-차원 벡터를 k-차원으로 매핑할 수 있는 알고리즘이다. 본 논문에서 FastMap을 매핑 알고리즘으로 사용한 이유는 두가지이다. 첫째, 대용량의 데이터베이스에서 MDS에 비해 FastMap는 빠른 검색과 업데이트가 가능하다. 둘째는 MDS는 소용량의 데이터베이스에서 매핑 알고리즘으로 적합하나 FastMap는 데이터베이스 용량에 관계 없이 임의의 오브젝트를 k-차원으로 매핑하는데 수행 복잡도는 $O(k)$이다.

또한, 벡터 공간 모델의 유사성 거리를 k-차원으로 매핑 하였을 때 일반적인 장점은 첫째 R*-트리와 같은 공간 접근 방법을 사용할 수 있어 질의에 대한 탐색 시간(search time)을 빠르게 한다. 둘째는 k-차원의 매핑 분포 모양으로 고차원 데이터집합의 구조를 밝혀 시각화와 클러스터 분석에 도움을 준다 [3][7][8].

3.5 사용자 인터페이스



[그림 4] 웹 문서간의 유사성을 2 차원 공간으로 매핑한 사용자 인터페이스

사용자 인터페이스(User Interface)는 사용자와 상호 작용으로 방문중인 웹 문서에 대한 정보를 보여준다.

<그림 4>는 2 차원 공간으로 매핑한 결과를 나타내고 있으며 현재 방문중인 웹 문서가 점 객체 형태로 매핑되고 있다. 매핑된 주변의 객체들을 확대해서 보면 현재 방문 중인 웹 문서하고 가장 유사성이 있는 문서에 대한 정보를 쉽게 얻을 수 있다.

4. 결론 및 향후 연구방향

본 논문은 웹 문서 히스토리 정보를 효과적으로 관리하여 과거에 방문한 웹 문서를 쉽게 찾을 수 있는 웹 문서 시각화 시스템을 구현 하였다. 웹 문서 시각화 시스템은 사용자를 대신하여 웹 문서를 분석하고, 사용자가 현재 방문 중인 웹 문서가 2-차원 공간 어느 부류에 문서인지를 확인 할 수 있다. 이로 인해 사용자는 히스토리 정보에 있는 웹 문서의 자료를 얻기 위해 URL 을 일일이 검색하는 등의 시간과 노력을 줄일 수 있게 된다.

향후 과제로서 시각화 구현에 있어 2-차원 공간으로 매핑하면서 생기는 오차, 즉 stress를 최소화 시키는 방법의 연구와 사용자가 실용적으로 사용할 수 있는 사용자 인터페이스 구현에 대한 연구가 필요하다.

참고문헌

[1] Ron Weiss, Bienvenido Velesz, Mark A. Sheldon, "HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering", Programming Systems Research Group MIT Laboratory for Computer Science, MA 02139.

[2] 김진호, 류근호, "정보검색", pp 155 – 216, pp528 – 545, 시그마프레스, 1995.7.

[3] Christos Faloutsos and King-Ip (David) Lin, "FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets" ACM SIGMOD, May 1995, San Jose, CA, pp. 163-174. Also available as technical report.

[4] MSHTML Reference "http://msdn.microsoft.com/workshop/browser/mshtml/reference/reference.asp".

[5] 백혜정, 박영택, 윤석환, "사용자 관심도를 이용한 웹 에이전트", 정보처리학회지 4 권, 1997. 9.

[6] 이상섭, 소영준, 박영택, "개인 웹 에이전트를 위한 사용자 프로파일 구축", 정보과학학회 논문지, 1998.

[7] Oren Zamir, Oren Etziono, "Web Document Clustering: A Feasibility Demonstration" In proceeding of the 21th International ACM SIG IR Conference on Research and Development in Information Retrieval, 1998.

[8] Michael.A.Berry, "Data Mining Techniques For Marketing, Sales, and Customer Support", pp 158 –215, Wiley Computer Publishing, 1995.