의미 구역에 기반한 관련 웹 페이지 요약 기법

이시은[0] 황인준
아주대학교 정보통신전문대학원 정보통신공학과
{sfmovie, ehwang}@ajou.ac.kr

Summarizing relevant web pages based on semantic region

Sieun Lee[0] Eenjun Hwang
The Graduate School of Information and Communication, Ajou University

요 약

웹 상의 정보는 여러 페이지들에 걸쳐 표현되고 있으나 대부분의 웹 브라우저는 웹 페이지 단위로 정보를 다루고 있기 때문에 사용자는 원하는 정보를 얻기 위해 여러 웹 페이지들을 방문해야 한다. 본 논문에서는 사용자의 요구에 부합되는 정보를 검색해 여러 페이지 상에 흩어져 있는 정보들에 대해 쉽게 이해 할 수 있도록 컬렉션 페이지를 제공한다. 컬렉션 페이지는 검색 된 웹 페이지들의 링크 관계를 제공하여 페이지들 사이에서의 정보의 구성을 알 수 있게 하고, 관련도 높은 페이지들의 주요 내용을 미리 가져와 보여줌으로써 정보에 대한 접근성을 높인다. 이를 위해 페이지 안에서 시각적으로 구분되는 동일한 주제의 정보를 담은 블록을 의미 구역으로 정의하고 웹 페이지를 실제 정보의 단위인 의미 구역으로 나누었다. 또한 의미 구역단위의 검색으로 여러 주제의 정보를 담고 있는 웹 페이지에 대한 검색 결과의 정확성을 높인다.

1. 서 론

인터넷은 정보전달과 서비스 이용의 수단으로써 널리 사용되고 있다. 기존의 웹 브라우저는 웹 페이지를 독립적인 정보의 단위로 다루고 있기 때문에 사용자는 여러 웹 페이지 상에 흩어져 있는 정보에 빠르게 접근하지 못하고 있다.

웹 상에서 원하는 정보에 접근하기 까지는 다음과 같은 어려움이 따른다. (i) 웹 상의 정보가 하이퍼링크와 프레임을 이용해 여러 페이지에 걸쳐 표현되어 있기 때문에 필요한 정보를 얻기 위해 여러 페이지를 일일이 열어보아야 한다. (ii) 링크 텍스트나 이미지, 또는 ALT 태그를 통해 해당 페이지의 내용을 알 수 있지만, 전체 페이지의 내용을 다 대변하지 못하고 요약적인 수준의 정보만을 제공하기 때문에 의도하지 않은 정보를 담은 페이지로의 방문으로 시간을 소비하게 된다. (iii) 구성을 잘 알지 못하는 웹 사이트 내에서 원하는 정보를 찾기 쉽지 않다. 웹 사이트에서 사이트 맵을 제공하지만 복잡한 사이트 내의 간략한 구조만을 담고 있어 원하는 정보를 찾기 어렵다. 또 사이트 내의 검색 폼은 단순히 검색 결과 페이지들의 링크와 몇 줄의 상응하는 문장을 보여주는데 그치고 있어 결과 페이지들의 관계를 알기 힘들며, 결과 페이지들의 내용 또한 파악하기 어려운 점이 있다.

이런 문제들을 해결하기 위하여 본 논문에서는 사용자의 질의를 바탕으로 웹 사이트와 같이 일련의 링크들로 연결 된 페이지들을 검색한다. 그리고 질의에 부합되는 정보를 담은 페이지들의 관계를 찾아 제공하고, 관련된 페이지들의 가장 관련 깊은 의미 구역들을 이용하여 컬렉션(collection) 페이지를 생성하는 기법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 웹 정보의 접근성 향상을 위한 관련 연구에 대해 소개한다. 그리고 3장에서는 의미 구역 기반의 검색기법과 검색 결과 페이지들을 이용하여 컬렉션 페이지를 구성하는 과정에 대해서 논한다. 4장에서는 제안하는 기법의 성능을 평가하고 마지막으로 5장에서는 결론을 맺는다.

2. 관련 연구

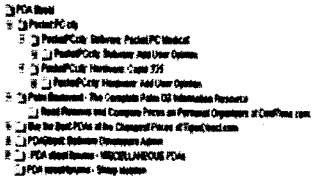
웹 페이지를 정보의 단위로 다루던 기존의 웹 정보 검색의 방식에서 벗어나 실제 정보의 단위를 정의하고 이를 웹 정보 검색에 이용해 검색의 효율을 높이는 연구가 진행되어왔다. 여러 페이지에 걸쳐 정보를 표현하는 경향을 반영하여, [1]은 검색 시 관련 된 웹 페이지들의 묶음을 하나의 논리적인 정보의 단위로 보고 이 연관 된 정보를 담은 페이지 셋(set)을 검색한다. 이를 통해 물리적인 페이지 단위의 정보가 아닌 논리적으로 연관 된 정보들을 함께 볼 수 있는 장점이 있다. MIU[2]는 검색 시 하나의 웹 페이지에 여러 주제의 정보가 함께 담길 수 있음을 고려해 페이지 내의 의미 구역을 검색에 활용하였다. 이는 사용자의 질의어 분포가 하나의 의미 구역에 밀집할수록 페이지에 포함 된 질의어들이 같은 주제의 정보를 나타낼 가능성이 높음을 고려한 것이다. 프레임과 테이블 구조를 이용해 한 페이지 안에 많은 정보를 담는 경향을 반영한 효율적인 검색 방법을 제시하였다.

링크 된 페이지의 내용을 미리 제공하여 사용자의 웹 탐색을 돕고 탐색 방향을 추천해주는 시스템에 대한 연구 또한 많이 이루어지고 있다. MS WebScout[3]은 마우스 포인터가 링크 텍스트 위에 위치했을 때 링크 페이지를 이미지 형태로 미리 가져와 보여준다. [4]는 링크 페이지의 저자, 제목, 언어, 그리고 접근 시간등의 정보를 제공한다. 또 사용자의 로그 정보를 마이닝하여 사용자의 브라우징 패턴과 경향을 파악해 사용자에게 유용한 링크, 사이트, 페이지 주소 등을 추천하는 연구 또한 많이 진행되고 있다. Hunter Gatherer[5]는 여러 페이지의 정보들을 빠르게 접근하고, 파악할 수 있도록 웹 페이지들의 컨텐츠를 모아 하나의 문서를 만들 수 있도록 돕는다. Web Skimming[6]은 웹 사이트 내의 페이지들 중 사용자 질의와 관련된 주요 문서들의 구성 순서(context path)를 찾아 제공한다.

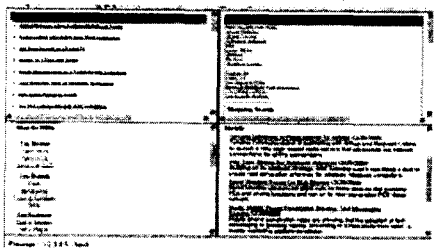
3. 컬렉션 페이지 생성

웹 사이트는 많은 웹 페이지들을 포함하고 있기 때문에 이

중에서 원하는 정보를 찾기가 쉽지 않다. 웹 사이트에서 검색 폼을 제공하긴 하지만 정보를 담은 페이지들의 관계와 사이트 내에서의 정보의 구성에 대해 이해하기 어렵고, 검색 결과 페이지들을 방문하지 않고는 페이지의 컨텐츠를 파악하기 어렵다. 본 논문에서는 사용자의 질의를 이용하여 웹 사이트 내의 페이지들을 검색한다. 그리고 질의에 부합되는 정보를 담은 페이지들의 순위를 매기고 이들의 링크 관계를 추출한다. 또한 검색 결과 페이지들의 가장 관련 깊은 의미구역들을 이용해 컬렉션 페이지를 생성한다.



(a) 결과 페이지들의 링크 트리



(b) 상위 랭크 된 페이지들의 주요 의미구역의 컬렉션

그림1. 컬렉션 페이지의 예

컬렉션 페이지는 검색된 웹 페이지들의 링크 관계를 트리 형태로 제공한다. 그림 1의 (a)는 결과 페이지들의 링크 트리를 보여주고 있다. 트리의 각 노드는 검색 된 페이지로 이동할 수 있는 링크를 담고 있으며 페이지의 컨텐츠에 대한 이해를 돕기 위해 각 페이지의 제목, 앵커(Anchor) 텍스트 등의 정보를 담고 있다. 또한 검색된 페이지들 중 상위에 랭크 된 페이지들의 가장 관련이 깊은 의미 구역을 보여줌으로써 페이지에 대한 이해를 높인다. 그림 1의 (b)는 컬렉션 페이지의 예를 보여준다. 이러한 컬렉션 페이지를 통해 사용자는 보다 쉽게 웹 사이트와 같은 페이지 셋(set)으로부터 원하는 정보를 찾을 수 있고, 하나의 페이지를 통해 여러 페이지의 내용을 동시에 볼 수 있다. 컬렉션 페이지 생성을 위해 우리는 페이지의 의미 구역을 추출하고 이를 바탕으로 제안하는 검색 알고리즘을 이용해 가장 관련 깊은 의미 구역을 추출한다.

3.1. 의미 구역 추출

HTML페이지는 사용자에게 이해하기 쉽고 보기 좋은 구조를 제공하기 위해 <table> 태그를 이용해 여러 겹의 중첩 테이블 구조를 갖는다. 이 테이블 구조는 계층적 특성 때문에 트리 구조로 표현될 수 있고, <table>, </table> 태그로 구분되는 웹 페이지의 컨텐츠는 이 트리의 노드들로 나누어진다. 많은 노드들의 컨텐츠는 독립적으로 정보를 전달하기에 너무 적은 내용을 가졌고, 단지 페이지의 구조를 형성하기 위해 쓰이고 있는 노드 또한 많기 때문에 주변 노드들을 병합하여 의미 구역을 형성한다. 관련 된 내용이 주변에 있을 가능성이 높기 때문에 주변 노드와 병합하는 것이다. 이 과정은 병합된 노드의 컨텐츠 양과 전체 페이지의 컨텐츠 양의 비가 임계값 이상이 될 때까지 계속된다. 이때 주변 노드들 중 형제 노드와 자식 노드가 병합 될 노드와 일관 된 정보를 담고 있다고 가정하고 주변 노드 중 형제 또는 자식 노드와 병합하여 의미 구역을 추출한다.

3.2. 의미구역 기반의 검색

컬렉션 페이지는 결과 페이지들의 의미 구역들 중 질의와 가장 관련 깊은 의미 구역들을 이용해 형성된다. 이를 위해 페이지 내 의미 구역들과 질의와의 유사도(similarity)를 계산하고 의미 구역들과 질의와의 유사도를 이용해 페이지의 점수가 계산 된다.

웹 검색을 위한 사용자 질의는 비교적 짧기 때문에 사용자가 요구한 질의어가 모두 등장하는 것이 질의어의 빈도수보다 중요하므로 [7] 이를 반영하여 식 1 과 같이 페이지 내의 각 의미 구역과 질의 간의 유사도를 계산한다. R은 의미 구역의 질의와의 유사도이고, N은 의미 구역 내에 등장한 질의어의 개수이고 F는 의미 구역 안에서 각 질의어의 등장 빈도수 중 가장 작은 빈도수 값이다. 이때 해당 의미 구역이 페이지의 메인 컨텐츠를 담은 구역이면 가중치를 준다. 메인 컨텐츠 구역은 페이지의 의미 구역들 중 가장 많은 컨텐츠를 가지고, 가장 적은 수의 링크를 가진 구역을 추출하여 얻어진다.

$$R = N*2 + F/2 \qquad (1)$$

페이지와 질의 간의 유사도는 의미 구역들의 유사도를 이용해 계산된다. 가장 높은 유사도의 의미 구역이 가장 중요한 구역이므로 이 구역의 유사도와 나머지 의미 구역의 유사도의 평균을 더해 페이지의 랭킹 점수를 얻게 된다. 식 2에서 Sim(q,p)는 사용자 질의와 페이지의 유사도를 뜻한다.

$$Sim(q,p) = Max(R) + \frac{\sum R_i}{Num(R_i)} \qquad (2)$$

또한 링크 페이지의 앵커 텍스트가 사용자의 질의어를 포함한다면, 이 페이지는 질의와 관련 가능성이 높다. 따라서 이러한 페이지의 경우 페이지의 유사도에 가중치를 준다. 그림 2는 링크 검색을 통해 의미 구역과 페이지의 유사도를 계산하는 알고리즘을 보여준다.

```
Function : MeasureSimilarity()
Parameter : target web page URL

Parse the target page
Look for the query term in each semantic region
Find the minimum frequency among frequencies of query terms for each
     semantic region
Find the main semantic region

for(i=0;i<totalRegionNum;i++)
    for(j=0;j<totalQueryNum;j++)
        if(frequency for query j in region I is higher than 0)
            count the number of query term occurred in a region

for(i=0;i<totalRegionNum;i++)
    calculate relevance score of each region
   // the number of occurring query term*2 + minimum frequency/2.0
   // if the query is composed only one query term,
   // the equation is occurring query terms+ minimum frequency/2.0

   if(region is the main region)
      relevance score of region=relevance score of region*1.2
   // give a weight the region that contains main region

Find the semantic region whose relevance score is highest
// most matched region


calculate the page relevance
    //the score for most matched region +
    //sum of other region's score/(totalRegionNum-1)

if(linking text for this page contains query term)
    relevance score of page = relevance score of page*1.5
```

그림 2. 웹 페이지와 의미구역의 사용자 질의 간의 유사도 측정

3.3. 시스템 구조

그림 4는 컬렉션 페이지 생성을 위한 시스템의 전체적인 구조를 보여준다. 전체 시스템은 크게 페이지 분석기(Page

Analyzer), 서비스 처리기(Service Handler), 페이지 생성기(Transcoder)로 나뉜다. 검색하는 각 페이지에 대해 페이지 분석기가 의미 구역을 추출하고, 서비스 처리기는 추출된 의미구역과 질의와의 유사도를 계산한 뒤 방문한 페이지들의 링크 관계 정보를 저장한다. 이러한 정보들을 바탕으로 페이지 생성기는 컬렉션 페이지를 생성한다.

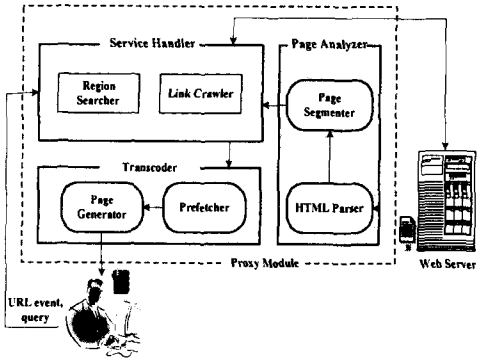


그림 4. 컬렉션 페이지 생성 시스템

컬렉션 페이지 생성을 위해 사용자가 지정한 한 웹 페이지로부터 시작하여 링크 된 페이지들을 검색하며 각 페이지의 의미구역의 유사도를 계산한다. 이때 각 페이지의 모든 링크가 유용한 정보를 담고 있지는 않기 때문에, 관련 없는 페이지들로의 불필요한 방문을 피하기 위해 다음과 같은 몇 가지 정책을 바탕으로 각 링크 페이지들을 검색한다.

(i) 최대 검색할 링크들의 깊이(depth)를 사용자가 지정한다. 이 범위 내의 링크 페이지들에 대해서만 검색한다.

(ii) 링크가 위치한 의미 구역의 유사도를 바탕으로 해당 링크 페이지로부터의 검색 깊이가 동적으로 계산된다. 링크가 위치한 의미 구역이 질의와 관련 깊은 구역이면 이 링크 페이지로부터의 검색 깊이는 증가시킨다. 따라서 관련 없는 의미 구역의 링크 페이지로부터의 검색보다 많은 깊이의 링크들을 따라가며 검색한다.

(iii) 링크 페이지가 검색이 시작 된 페이지와 다른 사이트의 페이지이고 질의어가 나타나지 않으면 이 페이지로부터의 링크들은 검색하지 않는다.

4. 실험 및 분석

컬렉션 페이지 생성 시스템의 성능을 평가하기 위해 웹 페이지를 의미 구역으로 나누고 이들 중 페이지의 메인 컨텐츠를 갖는 의미 구역을 추출하여 실제 메인 컨텐츠와의 비교를 통해 의미 구역 추출의 정확도(precision)를 측정하였다. 이를 위해 메인 컨텐츠를 구별할 수 있는 페이지들을 선정하였다. 표 1은 실험 대상이 된 페이지들의 메인 컨텐츠를 보여준다. 정확도는 사용자에 의해 인식 된 메인 컨텐츠 구역과 시스템에 의해 추출되는 구역의 유사도를 통하여 얻어진다. 그림 5는 추출 된 메인 컨텐츠의 정확도를 보여준다. 평균 정확도 94%로 메인 컨텐츠 추출이 이루어졌다.

추출된 메인 컨텐츠 구역은 웹 페이지들의 내용을 쉽게 파악할 수 있게 한다. 따라서 컬렉션 페이지 내의 여러 웹 페이지들의 내용을 동시에 볼 수 있도록 하여 링크들을 탐색하는 시간을 줄인다.

표 1. Experiment web page

Site	Main Contents	URL
Yahoo	**Weather**: Current condition, Local Forest	http://weather.yahoo.com/forecast/KSXX0037_f.html
	Stock: Market Summary	http://finance.yahoo.com/?u
CNN	**World** news: S. Korea to send envoy to Pyongyang	http://www.cnn.com/2003/WORLD/asiapcf/east/01/24/koreas.talks/index.html
	Tech. news: Nintendo to launch snazzier console	http://www.cnn.com/2003/TECH/fun.games/01/23/console.nintendo.reut/index.html
White House	**News** & Policies: A Column by Dr. Condoleezza Rice	http://www.whitehouse.gov/news/releases/2003/01/20030123-1.html
	Education: President Bush Celebrates First Anniversary of No Child Left Behind	http://www.whitehouse.gov/news/releases/2003/01/20030108-4.html

5. 결론

본 논문에서는 웹 접근성 향상을 위해 한 페이지로부터 시작되는 일련의 링크 페이지들을 검색하여 관련 된 정보들을 한 페이지에서 볼 수 있도록 컬렉션 페이지를 생성하는 과정을 살펴보았다. 본 시스템은 기존의 웹 검색엔진과 웹 브라우저가 하나의 웹 페이지 단위로 검색하고 정보를 제공하는 한계를 벗어나, 여러 페이지에 흩어져 있는 정보들을 하나의 페이지에 정리하여 표현해줌으로써 보다 빠르고 쉽게 페이지들 내의 정보의 구성과 각 페이지의 내용을 이해할 수 있도록 하였다.

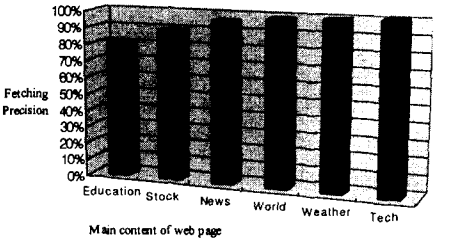


그림 5. 메인 컨텐츠 추출의 정확도

6. 참고 문헌

[1] Wen-Syan Li, K. Selcuk C., and Quoc V., Divyakant A., "Retrieving and Organizing Web Pages by Information Unit," 10th Int'l WWW Conference, 2001.

[2] Xiaoli Li, Bing L., Tong-Heng P., and Minqing H., "Web Search Based on Micro Information Units," 11th Int'l WWW Conference, 2002.

[3] Natasa M., Ralph S., and Robert T., "MS WebScout: Web Navigation Aid and Personal Web History Explorer," 11th Int'l WWW Conference, 2002.

[4] Ramesh R. Sarukkai, "Link Prediction and Path Analysis Using Markov Chains," 9th Int'l WWW Conference, 2000.

[5] M.C. Schraefel, et al., "Hunter Gatherer: Integration Support for the Creation and Management of Within-Web-Page Collections", 11th Int'l WWW Conference 2002.

[6] Kazutoshi S, et al., "Web Skimming: An Automatic Navigation Method along Context-path for Web Documents," 11th Int'l WWW Conference, 2002.

[7] Vo N. A., et al., "Impact Transformation: Effective and Efficient Web Retrieval", 25th Int'l ACM SIGIR Conference, 2002.