학습을 통한 주제기반 모바일 웹 콘텐츠 적응화

이은실° 강진범 양재영 최중민
지능시스템 연구실, 한양대학교
{eslee°, jbkang, jyyang, jmchoi}@cse.hanyang.ac.kr

Topic-Specific Mobile Web Content Adaptation through Learning

Eunshil Lee° Jinbeom Kang, Jaeyoung Yang, Chongmin Choi
Intelligent Systems Laboratory, Hanyang University

요 약

본 논문에서는 시각적 웹페이지 세그멘테이션 기법을 웹 콘텐츠 변환에 적용하고 이를 사용하여 이동기기에 적합한 개인화 기법을 제안한다. 웹페이지를 사람이 시각적으로 구분하는 것과 유사한 블록으로 나누고, 각 블록의 속성을 파악하여 불필요한 블록은 필터링한다 그리고 실제 내용을 나타내는 블록의 주제를 추출하여 휴대장치에 제공하는 효율적인 콘텐츠 적응화 기법을 제시한다. 또한 개인화된 콘텐츠를 제공하기 위해 적응화 과정에서 학습을 기반으로 사용자가 선호하는 정보만을 제공할 수 있는 개인화 기법을 제시한다.

1. 서 론

휴대전화나 PDA와 같은 다양한 휴대용 기기들을 사용하는 사용자들은 PC에서 제공받았던 풍부하고 다양한 콘텐츠를 이동형 휴대장치에서도 받아 볼 수 있기를 원하지만 기기마다 다른 성능과 사용자 선호도, 네트워크 대여폭 때문에 서비스는 제한될 수밖에 없다. 따라서 사용자가 사용중인 기기에 맞게 콘텐츠를 적응화(Adaptation)하는 과정이 필요하다. PC에서 제공되는 콘텐츠는 대부분 이동형 기기에 바로 표시되기에는 부적절한 경우가 대부분이다. 따라서 이렇게 상대적으로 대용량인 콘텐츠를 기기에 맞게 효과적으로 보여주는 방법에 대한 연구가 최근 이슈가 되고 있다.

본 논문에서는 시각적 웹페이지 세그멘테이션 기법을 웹 콘텐츠 변환에 적용하고 이를 사용하여 이동기기에 적합한 개인화 기법을 제안한다. 즉, 웹페이지를 사람이 시각적으로 구분하는 것과 유사한 블록으로 나누고, 각 블록의 속성을 파악하여 불필요한 블록은 필터링한다. 실제 내용을 나타내는 블록의 주제를 추출하여 휴대장치에 제공하는 효율적인 콘텐츠 적응화 기법을 제시한다. 또한 개인화된 콘텐츠를 제공하기 위해 적응화 과정에서 학습을 기반으로 사용자가 선호하는 정보만을 빠르게 제공할 수 있는 개인화 기법을 제시한다.

2. 관련연구

모바일 콘텐츠 적응화 연구는 2001년을 기준으로 나누어 볼 수 있다. 기준 이전 연구에서는 페이지의 URL을 기기에 입력하면 프록시 서버가 사용자 원하는 페이지의 내용 중에서 하이퍼링크로 표현된 정보를 중심으로 사용자의 기기에 보여주는 형태를 취했다.[1][2]

2001년 이후 연구부터는 시각적 웹페이지 세그멘테이션 기법들이 제안되면서 웹페이지의 정보를 오브젝트나 블록단위로 나누게 되었다. 블록단위로 콘텐츠를 이동기기에 제공하여 이전의 방식과는 다르게 링크 주위의 정보도 보여줄 수 있어서 효율적인 정보제공이 가능해졌다.[3][4][5]

링크 중심의 방식은 링크 위주의 콘텐츠를 제공하므로 정보를 폭넓게 전하는 방법에 문제가 발생할 수 있다. 그러나 블록 단위의 세그멘테이션 기법을 이용하면 링크 주변 정보를 추출할 수 있어서 다양한 정보를 사용자에게 전달할 수 있는 장점이 있다.

3. 기존 적응화 방법의 문제점과 해결방안

기존의 적응화 방법은 페이지의 하이퍼링크로 표현된 정보만을 추출해 이를 중심으로 사용자의 휴대장치에 정보를 제공한다. 이러한 방법을 사용하면 링크 주위의 정보를 알 수 없어 사용자가 원하는 정보인지 판단하기 힘들다. 또 정보를 얻기 위해 몇 번의 하이퍼링크를 지나야 할지 알 수 없고 불필요한 페이지로 이어질 수도 있다. 위와 같은 방법은 사용자에게 정보에 대한 불신감을 안겨줄 수도 있게 된다.

이러한 점을 해결하기 위해 본 논문에서는 링크중심이 아닌 블록중심의 적응화 방법을 사용하고 있다. 웹페이지를 블록단위로 나눠서 네비게이션 바, 네비게이션 리스트, 콘텐츠와 같은 카테고리 블록으로 구분한다. 카테고리를 분류해서 어떤 정보를 이용하고 있는지 알 수 있다. 또 블록단위로 제목과 요약한 내용을 제공함으로서 원하는 정보가 아닌 경우 접근하지 않고 원하는 정보를 찾기 위한 시간을 절약할 수 있다. 본 논문에서는 위와 같은 방법 이외에도 사용자가 블록에 접근하여 상세 페이지를 선택하면 블록내의 최대 빈도 단어가 저장되고 이 정보를 통해 사용자가 원하는 정보에 대한 우선순위가 높아진다. 본 논문에서는 링크중심방식에서 알 수 없었던 링크 주위의 정보가 무엇을 나타내는지 사용자에게 전해줌으로서 사용자의 확실한 정보판단에 도움을 준다.

4. 제안하는 방법

본 논문에서 제안하는 모바일 적응화 과정은 그림 1에서와 같은 구조로 되어 있다. 우선, 시각적인 웹 페이지 세그멘테이션 기법인 VIPS[4]를 이용하여 블록을 추출한다. 추출된 블록을 통하여 광고나 카피라이터 등을 나타내는 필요 없는 블록을 제거한다. 남은 블록에서 제목과 요약된 내용을 사용자의 디바이스에 맞게 변형시켜 제공한다. 사용자는 디바이스를 통해 원하는 정보를 획득한다. 사용자가 이용한 콘텐츠에서 최대 빈도 단어를 추출하여 데이터베이스에 추가하여 추가된 단어들을 바탕으로 블록의 우선순위를 변경시켜 제공하게 된다.

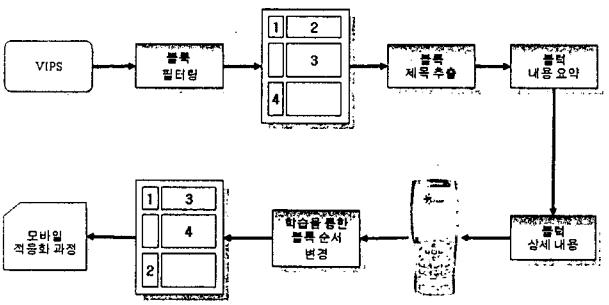

그림 1 적응화 처리과정

4.1. 블록 필터링

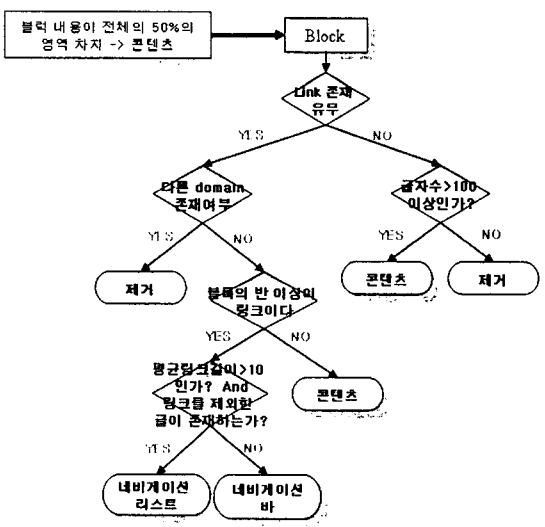

그림 2 블록 카테고리 분류 규칙

블록 필터링 단계는 규칙을 이용해서 VIPS를 통해 나누어진 블록들을 세가지 블록으로 나누고 불필요한 블록은 제거하는 과정이다.
블록은 콘텐츠, 네비게이션 바, 네비게이션 리스트와 같은 세가지 카테고리로 나눈다.

- 콘텐츠 : 내용을 가지고 있는 블록
- 네비게이션 바 : 메뉴
- 네비게이션 리스트 : 내용을 포함할 가능성을 가지고 있는 링크 리스트

그림 2는 세가지 카테고리로 분류하기 위한 규칙이다.

4.2. 블록 제목 추출

블록 필터링을 통해 나누어진 블록 중 콘텐츠 블록에서 제목을 추출한다. 콘텐츠 블록의 내용을 토큰 단위로 몇 번 존재하는가를 카운트한다. 2번 이상 존재한다면 그 다음 토큰을 불러와서 다시 카운트를 하게 되고 1번 존재하게 되면 구(phrase)가 출력되고 첫 번째 토큰이 제거된다. 다음 토큰을 불러와서 카운트한다. 이러한 과정을 반복해 빈도수가 높은 구가 제목으로 선택된다.

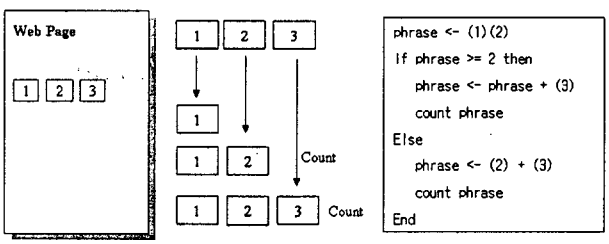

그림 3 제목 추출 알고리즘

4.3. 블록 내용 요약

블록 내용 요약은 4.2절의 과정에서 추출된 제목이 될 가능성이 있는 후보들을 사용한다. 후보 구들이 하나라도 존재하는 절을 추출한다. 이것을 요약후보라 하고 요약후보들은 예외처리과정을 거쳐 모바일 디바이스에서 보여지는 블록 내용 요약 문장이 된다.

4.4. 학습을 통한 개인화

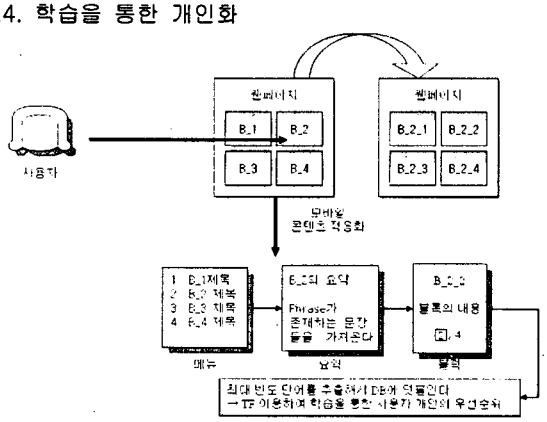

그림 4 학습을 통한 개인화 처리과정

제목추출과 내용요약 과정이 종료되면 디바이스에 정보가 제공된다. 사용자는 블록마다 추출된 제목을 보고 자신이 보고자 하는 내용을 선택하게 된다. 사용자가 내용

요약에서 상세정보를 보고자 선택하면 블록 내용에서 빈도수가 가장 높았던 단어가 추출되고 데이터베이스에 덧붙여진다. 정보검색의 벡터모델을 이용하여 블록내의 단어들과 학습에 의해 데이터베이스에 저장된 단어들을 벡터상에 표현한다. 데이터베이스에 저장된 단어들과의 코사인 측정값이 가장 적은 벡터상의 점으로 표현된 블록 순으로 블록의 우선순위를 변경한다.

5. 구현

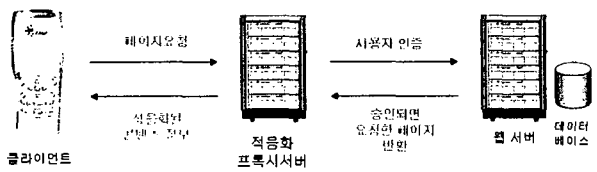

그림 5 시스템 구조

그림 5는 본 논문의 시스템 구조이다. 클라이언트-서버 구조를 취하고 있고 중간에 프록시 서버가 존재한다. 사용자는 디바이스에서 로그인을 하고 원하는 페이지의 주소를 입력한다. 프록시서버는 로그인 정보와 페이지 정보를 받아서 웹서버에 보내주고 웹 서버는 사용자 인증과정을 통해 요청한 페이지를 반환해 준다. 반환된 페이지는 프록시서버에서 그림 4에서 제안한 과정을 통해 모바일 디바이스에 맞게 적응화된 콘텐츠 정보를 사용자에게 보여준다.

학습과정에서는 사용자가 요약된 내용에서 상세내용보기를 선택하면 콘텐츠에서 빈도수가 가장 높은 단어가 프록시서버를 통해 기존에 저장된 데이터베이스에 덧붙여진다. 프록시서버는 데이터베이스에 있는 단어들과 사용자가 원하는 페이지 블록들을 벡터상에 표현하고 코사인값을 계산한다. 계산된 코사인값이 작은 블록 순으로 블록의 우선순위가 정해진다. 개인화에 따라 블록의 우선순위가 변경된 정보를 사용자 디바이스에 보여준다.

6. 성능평가

카테고리 분류	제목추출	내용요약	우선순위
77.5%	55%	67.5%	86.5%

표 2 성능평가 결과

표 2는 4가지 평가기준을 통해 나온 결과이다. CNN 사이트에서 200개의 페이지를 수집해서 실험했다. 결과 값은 한 문서상에 올바른 정보 대한 평가된 정보를 정확도로 나타내었다. 카테고리 분류, 블록의 내용요약, 블록의 우선순위는 60%를 넘었으나 제목추출은 60%이하의 성능을 보였다. 그 이유는 웹 사이트의 인덱스 페이지에서 콘텐츠 블록이라고 잘못 분류될 경우 제목추출이 제대로 이루어지지 않기 때문이다. 내용요약은 제목추출을 위한 후보 구를 모두 사용하므로 제목추출보다 좋은 성능을 보였다.

7. 결론

본 논문은 개인화된 콘텐츠 제공을 위해 시각적 웹 페이지 세그멘테이션 기법을 적용하여 웹페이지의 콘텐츠 블록을 구분하고, 콘텐츠 블록들의 주제를 추출, 내용을 요약한다. 이러한 적응화 과정을 기반으로 모바일 디바이스에 제공된 정보를 사용자가 이용패턴을 학습하여 사용자가 원하는 정보의 우선순위를 높여 휴대장치에 제공하는 효율적인 콘텐츠 적응화 기법을 제시했다.

링크중심의 적응화 방법은 하이퍼링크 주변의 내용을 사용자에게 알려주기 힘들었으나, 블록중심의 적응화 방법은 사용자에게 블록 자체의 내용을 전달해 줄 수 있어 전체적으로 어떤 내용을 전달하려고 하는지 파악할 수 있게 되었다.

또한 본 논문에서는 블록의 내용을 링크 위주로만 보는 것이 아니라 제목추출과 내용을 요약하는 링크를 기반으로 하는 내용 중심적인 방법을 제시하고 있다.

본 논문에서 제시한 방법을 이용한 적응화 실험에서 제목추출 결과가 가장 낮게 나왔다. 이유는 카테고리 분류시 몇 개의 블록이 컨텐츠 블록으로 잘못 분류됨으로서 측정값에 영향을 주었다.

향후 카테고리 분류 규칙과 제목추출 알고리즘을 개선한다면 더 좋은 성능이 예상된다.

참고문헌

[1] T. Laakko and T. Hiltunen, "Adapting Web Content to Mobile User Agents", IEEE Internet Computing Vol.9, No.2,pp46-53,March/April 2005.

[2] A. Pashtan, S. Kollipara, and M. Pearce, "Adapting Content for Wireless Web Services," IEEE Internet Computing, vol. 7, no. 5, 2003, pp. 79-85.

[3] Jinlin Chen, Baoyao Zhou, Jin Shi, Hongjiang Zhang, Oiu Fengwu, "Function-based Object Model Towards Website Adaptation", Proceedings of the 10th international conference on World Wide Web,2001.

[4] D.Cai, S. Yu, J. Wen, and W. Ma, "VIPS: A Vision-based Page Segmentation Algorithm", Microsoft Technical Report, MSR-TR-2003-79, 2003.

[5] Ruihua Song, Haifeng Liu, Ji-Rong Wen, Wei-Ying Ma, "Learning block importance models for web pages", Proceedings of the 13th international conference on World Wide Web, New York, USA, 2004.