Web 문서의 효율적인 실시간 검색을 위한 잡음 제거와 패턴 정합 기법

강대기, 이제선, 함호상
전자 통신 연구원 컴퓨터•소프트웨어 연구소
시스템 통합 연구부 전자 거래 연구팀

Noise Removal and Pattern Matching for Efficient Meta-Search of Web Documents

°Dae-Ki Kang, Jeseon Lee, Hosang Ham
ETRI Computer & Software Technology Laboratory
System Integration Department Electronic Commerce Team

요약

웹 상의 메타 검색 엔진, Push 프로그램 그리고 에이전트와 같은 웹 기반 응용 프로그램들은 웹 문서의 취득과 자동 튜터링에 대한 능력을 필요로 한다. 이를 위한 인터페이스의 지식들은 대부분 코드 내에서 ad-hoc 으로 구현되어 왔다. 본 논문에서는 취득된 웹 문서를 전처리하고 원하는 정보를 추출하기 위한 방법을 제시하고, 웹 상의 신문 기사에 대한 검색으로 실험해 보았다. 검색 시스템은 웹 문서의 전처리 과정을 통해 필요한 정보에만 집중할 수 있고, 아주 적은 양의 일반화된 지식을 토대로 원하는 정보를 용이하게 찾을 수 있었으며, 또한 웹 문서의 형식이 바뀌더라도 크게 영향을 받지 않으며, 새로운 웹 사이트의 추가도 용이하였다. 본 논문의 방법으로 구현된 신문 기사 검색 시스템은, URL 과 아주 적은 양의 지식만으로도, 10 개의 신문 웹 사이트에서 문서를 가져와 효과적으로 해석할 수 있었다. 본 논문의 방법은 메타 검색 엔진이나, 잡지나 신문 기사 정보의 푸쉬(Push) 솔루션, 또는 상품 정보 검색 시스템 등의 설계에 활용될 수 있다.

1. 서론

웹이 보편화되면서 수많은 정보들로 넘치게 되었다. 그러나 이렇게 많은 정보들로 인해 오히려 사용자들은 자신이 찾고자 하는 정보를 찾는 데 많은 시간을 들이게 되었다[1]. 이러한 사용자들의 요구를 바탕으로 검색 엔진들과 메타 검색 엔진들이 개발되어 서비스되고 있다. 그러나 보다 정확한 검색을 위해서 전문 분야에 대한 검색 엔진이 최근 들어 요구되고 있다. 이를테면 전자 상거래에서의 상품 정보 검색 엔진이나 쇼핑 에이전트들[2]이 그러한 예이다.

특히 신문이나 TV, 잡지의 인터넷 진출이 두드러지고 있다. 이러한 사이트들은 정해진 주기마다 내용을 바꾸고 있으며, 기본적으로 가지고 있는 다른 매체에서의 지명도를 토대로 웹에서의 영향력을 계속 높여가고 있다. 또한 이른바 푸쉬(Push) 솔루션들은 특히 이러한 언론 및 방송 매체들의 정보들을, 웹 브라우저를 가지고 사용자가 찾아가는 것이 아니라, TV 채널을 시청하는 것과 같이 볼 수 있도록 해준다.

본 논문에서는 이러한 쇼핑 에이전트들이나 Push 솔루션들이 실시간으로 웹 문서에서 원하는 정보를 검색하여 추출하는 과정을 보다 효율적으로 할 수 있게 하면서, 웹 문서의 유동적인 변화에 크게 영향을 받지 않게 하기 위한 잡음 제거 및 패턴 정합 방법을 제안하고, 이를 신문 기사 검색 시스템에 실험해 보았다.

본 논문의 구성은 다음과 같다. 2 장에서는 검색 에이전트 및 푸쉬 솔루션들에 관한 관련 연구 및 상용화되어 서비스 되고 있는 시스템들을 기술하고, 제 3 장에서는 잡음 제거 방법과 패턴 정합 방법을 제안하고, 제 4 장에서는 이러한 방법들을 10 개의 신문 사이트에 대한 실험한 결과를 기술하고, 마지막으로 제 5 장에서는 결론 및 향후 연구 과제를 기술한다.

2. 관련 연구

대부분의 정보 검색 시스템들은 웹 문서를 가져와서 형태소 분석 등을 통해 키워드들을 추출한 다음, 필요한 가중치 및 문서 내의 위치 정보를 계산하고

난 후 저장하는 접근 방법을 취하고 있다. 최근에 등장한 메타 검색 엔진들과 검색 에이전트들은 이러한 검색 엔진들의 결과를 이용하는 방법을 취하고 있다. 대표적인 메타 검색 엔진은 미스 다찾니, MetaCrawler 등이 있다. 이와 비슷한 서비스로 웹 브라우저를 가지고 사용자가 찾아가는 것이 아니라, TV 채널을 시청하는 것과 같이 볼 수 있도록 해 주는 푸쉬 솔루션들이 있다. 대표적인 푸쉬 솔루션은 PCN(PointCast Network), Netscape 의 NetCaster, Microsoft 의 MSNBC, Marimba 의 CasterNet, The DJ Web Radio, After Dark Online, NCK Telecom 의 IIC, NcoWiz 의 LiveCAST, HiCast 등이 있다. 이러한 메타 검색 시스템들이나 푸쉬 솔루션들은 조회할 검색 엔진에 대해 접근하고 웹 문서를 가지와서 원하는 내용을 추출하기 위한 방법을 필요로 한다.

이러한 인터페이스를 위한 지식은 대부분의 경우 사람의 손에 의해 프로그램 내에 코드로 반영되는 경우가 많다. 그러나 최근의 ShopBot 과 같은 쇼핑 에이전트[2]들은 조회할 검색 엔진의 웹 문서를 정규식으로 변환하여 학습할 수 있는 능력들을 가지고 있다

3. 잡음 제거 및 패턴 정합 방법

웹 문서는 내부에 많은 부가적인 태그(tag) 정보를 가지고 있다. 이러한 태그 정보들은 웹 문서를 보다 효과적으로 표현하기 위해 필요한 것들이다. 실제로 사용자가 얻는 정보는 태그 정보가 아닌 일반 텍스트 정보이다. 대부분의 정보 검색 엔진들은 태그 정보들을 URL 추출과 가중치 계산에만 사용할 뿐, 거의 남기지 않고 전부 없애는 경우가 많다.

그러나 실시간 검색이나 쇼핑 에이전트의 경우에는, 특정 키워드가 문서 내의 어느 부분에 위치하는가가 중요하므로, 태그 정보를 잡음과 그렇지 않은 정보로 분류하여 잡음 제거를 통해 잡음 정보만 없앤다. 일단 웹 문서에서 잡음 정보가 없어지면, 결과 문서는 정규식으로 변환된다. 변환된 정규식에서 원하는 정보의 패턴이 스트링 정합(string match) 방법으로 추출된다.

3.1 잡음 제거

일반 검색 엔진의 경우에는 <TITLE> 태그나 <H1>, <H2>, <B>, <META> 태그 등을 이용하여 키워드의 가중치를 높일 수 있다. 일반 검색 엔진들에서 중요한 것은 사용자의 입력된 키워드와 가장 가까운 단어를 찾는 것이다. 그러나 검색 엔진들에게 질의하여 결과를 추출하는 메타 검색이나 도서나 CD 의 실시간 검색과 같은 쇼핑 에이전트에서는 일반 검색과 같이 사용자의 요구를 미리 모르는 상황에서 문서를 색인하는 것이 아니라, 사용자의 요구를 이미 아는 상황에서의 검색 결과를 2 차적으로 필터링하고 원하는 정보만을 추출하는 것이 문제이다. 이를 위해서는 문서 내의 특정 태그들의 정보가 유지되어야 한다. 이러한 태그들은 <TABLE>, <TR>, <TD>, <LI>, <P>,
 등과 같이 문서 내에서 하나의 행(line)이나 레코드를 구성하기 위한 것들이다. 이러한 하나의 행이나 레코드를 구성하는 태그들이 중요한 이유는 대부분의 검색 결과가 테이블이나 리스트의 형태로 나오는 경우가 많기 때문이다. 웹을 통해 가져온 문서는 이러한 태그들과 텍스트 정보만 남겨지고 잡음 제거가 되는 데, 이러한 잡음 제거 모듈은 lex 나, awk, perl 등로 제작된 파서를 통해 쉽게 구현될 수 있다. <그림 1>은 본 논문의 신문 기사 검색 시스템에서 사용된 잡음 제거를 위한 lex 소스의 일부이다.

```
\<[tT][aA][bB][lL][eE][^\>]*[\>]                    {
  state++; if (!script_state) fprintf(yyout,"\n<TABLE>\n"); }
\<\/[tT][aA][bB][lL][eE][^\>]*[\>]    {
  state--; if (!script_state) fprintf(yyout,"\n</TABLE>\n"); }
\<[tT][rR][^\>]*\>          {
  if (!script_state) fprintf(yyout,"\n<TR>\n"); }
\<\/[tT][rR][^\>]*\>        {
  if (!script_state) fprintf(yyout,"\n</TR>\n"); }
\<[tT][hH]\>                {
  if (!script_state) fprintf(yyout,"\n<TH>\n"); }
\<[tT][hH][^\>]*\>          {
  if (!script_state) fprintf(yyout,"\n%s\n",toupperstr(yytext)); }
\<\/[tT][hH][^\>]*\>        {
  if (!script_state) fprintf(yyout,"\n</TH>\n"); }
\<[tT][dD]\>                {
  if (!script_state) fprintf(yyout,"\n<TD>\n"); }
\<[tT][dD][^\>]*\>          {
  if (!script_state) fprintf(yyout,"\n%s\n",toupperstr(yytext)); }
\<\/[tT][dD][^\>]*\>        {
  if (!script_state) fprintf(yyout,"\n</TD>\n"); }
```

그림 1 잡음 제거를 위한 lex 코드

3.2 패턴 정합

웹 문서에서 잡음 제거 과정을 거쳐, 정보 추출에 필요한 데이터들만 남은 결과 문서의 각각의 요소들은 정규식으로 변환된다. 변환된 정규식의 시퀀스에서 원하는 정보의 패턴이 스트링 정합(string match) 방법으로 추출될 수 있다. 원하는 정보의 패턴이 존재하는 이유는, 대부분의 실시간 검색 내지 메타 검색 응용에서 원하는 정보는 검색 엔진의 CGI(Common Gateway Interface) 프로그램에 의해 일정한 패턴 내에 등장하기 때문이다. 또한 원하는 정보의 패턴은 실제 응용 예에 따라 다르나, 대부분의 경우 단순한 가정이나 직관에서 얻어질 수 있다.

이해를 돕기 위해, 예를 들어 신문 기사 검색 시스템을 보면, 신문의 홈페이지에서 각각의 분류에 대한 페이지들이 존재하고, 각각의 분류에서 기사의 제목은 주로 <A> 태그에 의해 둘러싸여지게 된다.

<그림 2>는 신문 기사 검색 시스템에서, "전자 신문"의 "경제, 과학" 면을 가져와서 잡음 처리한 예이

다.

```
<TABLE>
<TD>
<A>
HREF="http://www.etnews.co.kr/etnews/new_etnews_content?199808220009102"
STEPI 보고서 "국가혁신체계 하부구조 취약"
</A>
국가경쟁력을 가름하는 기술혁신을 바탕으로 한 우리의 국가혁신체세가 취약한 것으로 드러났다
21 일 과학기술정책관리연구소(STEPI, 소장 장문호)가 continue
</TABLE>
```

그림 2 잡음 처리된 "전자 신문"의 "경제, 과학" 분류

<그림 2>의 행은 <그림 3>과 같은 스트링 치환을 통해 "TDAUXaXt"라는 스트링으로 변환될 수 있다.

```
<TABLE> →T
<D> →D
<A> →A
HREF= →U
일반 텍스트 →X
</A> →a
</TABLE> →t
```

그림 3 정규식 생성을 위한 스트링 치환

대부분의 기사 제목은 주로 하이퍼 링크 태그 안에 둘러싸이므로, 원하는 정보의 패턴은 "AUXa"로 나타낼 수 있으며, 위의 변환된 스트링 "TDAUXaXt"에서 검색될 수 있다. 이처럼 단순한 직관에서 나온 휴리스틱으로 간단히 신문 기사의 제목과 URL을 찾을 수 있음을 알 수 있다.

원하는 정보의 패턴은 별도의 학습 방식을 통해 학습될 수도 있다. 이 경우 패턴 추출을 위한 학습이 필요하게 된다. 본 논문에서는 이러한 학습 방법으로 주어진 문서에서 가장 많이 등장하는 레코드의 패턴에 대해 점수를 매기는 방법을 사용하였다. 예를 들면 신문 기사의 분류 페이지에서 가장 많이 등장하는 패턴은 "AUXa"로 기사의 제목과 그 URL이었다. 상품 정보 검색 에이전트에서 가장 많이 등장하는 패턴은 "DXd"로 테이블 데이터와 그 텍스트이다. 그러나 별도의 메타 검색이나 실시간 검색의 응용에 따라 이러한 학습 방법은 따로 고안되어야 할 것이다.

4. 실험 결과

본 논문에서는 이러한 잡음 제거와 패턴 정합 방법을 신문 기사 검색 시스템을 구현하여 실험해 보았다. 대부분의 신문 사이트는 "정치", "경제"와 같은 분류를 가지고 있고, 각 분류에서 각각의 기사는 하이퍼링크의 태그로 둘러싸여진 제목의 형태로 표현되어 있다.

<그림 4>는 신문 기사 검색 시스템에서 "문화 일보"의 "정치/국제" 면을 검색한 예이다.

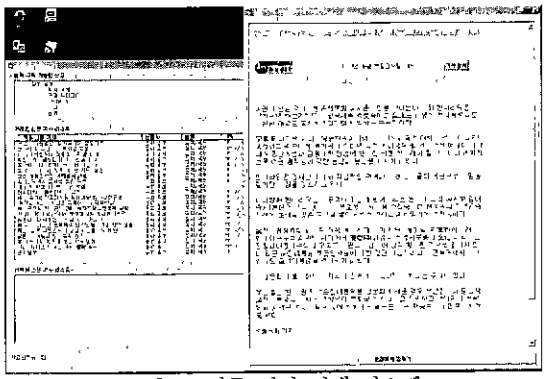

그림 4 신문 기사 검색 시스템

본 신문 기사 검색 시스템은 10 개의 신문 사이트에서 원하는 분류면을 가져와서 기사들과 그 URL을 추출한다. 사용되는 정보의 패턴은 "AUXa", "XAUa", 그리고 "AUaX" 이다.

신문 기사 검색 시스템은 Visual Basic 언어와 lex, C 언어로 제작되었으며, lex 와 C 언어로 짜여진 프로그램은 Visual C++ 컴파일러를 통해 DLL(Dynamic Link Library)로 제작되었으며, 검색된 신문 기사들 중 저장하기 위해 선택된 것들을 PDA(Personal Data Assistant)에 저장할 수 있는 기능도 가지고 있다.

5. 결론

인공 지능의 응용 중에서 영상 처리나 음성 처리의 예를 보면 너무 많은 데이터들 중 필요한 정보만을 추출하는 잡음 제거와 특징 벡터 추출과 같은 과정이 있다. 본 논문에서의 잡음 제거 과정 또한 적용 분야는 다르나, 불필요하게 중복된 정보를 제거한다는 점은 동일하다고 볼 수 있다.

본 신문 기사 검색 시스템은 기존의 Push 솔루션들과 비교해 볼 때, 신문 사이트의 URL과 분류 및 URL, 그리고 아주 적은 양의 패턴 정보를 지식으로 동일한 성능을 보이고 있다. 또한 새로운 사이트의 추가도 빠른 시간 내에 용이하게 이루어지는 장점들을 가지고 있다. 다만, 몇 가지 휴리스틱과 패턴 정보를 더 추가하여 머리말과 꼬릿말 정보의 제거를 수행하는 문제가 남아 있다. 또한 신문 사이트의 검색 엔진을 활용하는 방안과 자체 키워드 검색 엔진을 구성하는 문제들이 차후 과제로 남아 있다.

참고 문헌

[1] O. Etzioni, "The World-Wide Web: Quagmire or Gold Mine," Communications of ACM, Vol. 39, No 11, pp. 1-5, Nov. 1996

[2] R. B Doorenbos, O. Etzioni, and D. S. Weld, "A Scalable Comparison-Shopping Agent for the World-Wide Web," *1st International Conference on Autonomous Agent*, Jan. 1997.