



PLASSE
RESEARCH

ROSAECcenter
Research On Software Analysis for Error-free Computing
소프트웨어 무결점 연구센터 KOSEF ERC

제7회 소프트웨어무결점연구센터 워크샵
@HKUST

요약파싱기법으로 동적으로 생성되는 HTML의 유효성 검사하기

김현하, 도경구 @ 한양대학교 PLASSE Research

David A. Schmidt @ Kansas State University

2012.1.19

순서

- ◆ 요약파싱기법
- ◆ HTML 유효성검사
- ◆ 진행과정
- ◆ 더 해야할 것

요약파싱기법으로 동적으로 생성되는 HTML의 유효성 검사하기

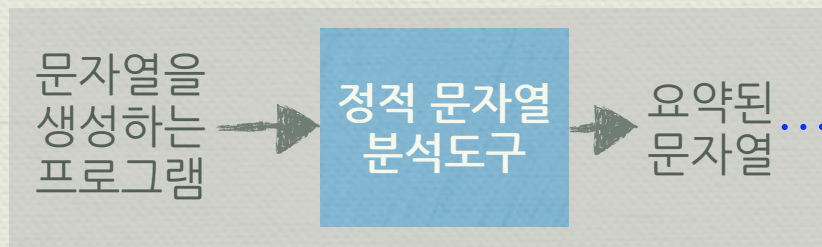
- ◆ 요약파싱기법
- ◆ HTML 유효성검사
- ◆ 진행과정
- ◆ 더 해야할 것

분석한 다음에 파싱

◆ 요약도메인이 정규표현식인 문자열 분석

요약해석기법으로 분석한 요약된 문자열이

바른 모양을 가지고 있는가?



Java프로그램이 생성하는 SQL 구문이

바른 문법을 가지는가?

어떤 질의들을 요청하는가?

어떤 테이블들을 접근하는가?

요약파싱

요약해석기법을 사용한 정적 문자열 분석도구

문자열을
생성하는
프로그램



정적 문자열
분석도구



요약된
문자열

+

문자열이 바른 문법인지 검사하는 파서

문자열



파서



바른문법



틀린문법

요약파싱의 요약도메인

파스 스택

- (1) $E \rightarrow E * B$
- (2) $E \rightarrow E + B$
- (3) $E \rightarrow B$
- (4) $B \rightarrow 0$
- (5) $B \rightarrow 1$

문법

state	action'					goto	
	*	+	0	1	\$	E	B
0			s1	s2		3	4
1	r4	r4	r4	r4	r4		
2	r5	r5	r5	r5	r5		
3	s5	s6			acc		
4	r3	r3	r3	r3	r3		
5			s1	s2			7
6			s1	s2			8
7	r1	r1	r1	r1	r1		
8	r2	r2	r2	r2	r2		

파싱 테이블

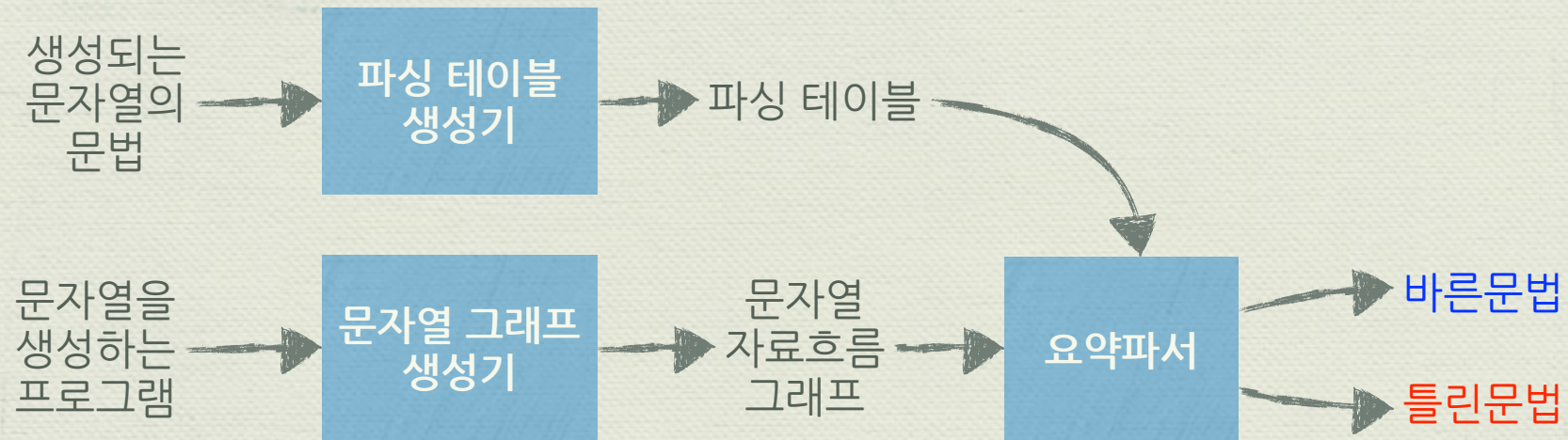
입력 : 1 + 1

State	Input Stream	Output Stream	Stack	Next Action
0	1+1\$		[0]	Shift 2
2	+1\$		[0,2]	Reduce 5
4	+1\$	5	[0,4]	Reduce 3
3	+1\$	5,3	[0,3]	Shift 6
6	1\$	5,3	[0,3,6]	Shift 2
2	\$	5,3	[0,3,6,2]	Reduce 5
8	\$	5,3,5	[0,3,6,8]	Reduce 2
3	\$	5,3,5,2	[0,3]	Accept

파싱결과

첫번째 결과물

- ◆ K.-G. Doh, H. Kim, and D.A. Schmidt, **Abstract parsing: static analysis of dynamically generated string output using LR-parsing technology**, SAS 2009



PHP프로그램이 생성하는 (X)HTML 문서가 바른 문법을 가지고 있는가?

요약파싱기법으로 동적으로 생성되는 HTML의 유효성 검사하기

- ◆ 요약파싱기법
- ◆ **HTML 유효성검사**
- ◆ 진행과정
- ◆ 더 해야할 것

HTML 유효성검사

- ◆ 웹 문서가 바른 모양을 가지고 있는지 검사
- ◆ 여러 브라우저에서 같은 모양으로 보기 위함
- ◆ 대표적인 유효성검사도구: W3C 마크업 밸리데이터

The screenshot shows the W3C Markup Validation Service interface. At the top, there's a blue header with the W3C logo and the text 'Markup Validation Service' and 'Check the markup (HTML, XHTML, ...) of Web documents'. Below this, there are three tabs: 'Validate by URI' (selected), 'Validate by File Upload', and 'Validate by Direct Input'. Under the 'Validate by URI' tab, there's a section titled 'Validate by URI' with the text 'Validate a document online:'. Below this, there's a label 'Address:' followed by a text input field. Under the input field, there's a link 'More Options'. At the bottom of the form, there's a 'Check' button. Below the form, there's a paragraph of text: 'This validator checks the [markup validity](#) of Web documents in HTML, XHTML, SMIL, MathML, etc. If you wish to validate specific content such as [RSS/Atom feeds](#) or [CSS stylesheets](#), [MobileOK content](#), or to [find broken links](#), there are [other validators and tools](#) available.'

<http://validator.w3.org>

HTML 유효성검사

◆ 브라우저마다 다르게 해석되는 페이지

```
<p><strong>This text should be bold</p>
```

```
<p>Should this text be bold? How does the HTML look when rendered?</p>
```

```
<p><a href="#"></strong>This text should be a link</p>
```

```
<p>Should this text be a link? How does the HTML look when rendered?</p>
```

출처: <http://thinkvitamin.com/code/same-dom-errors-different-browser-interpretations/>

유효하지
않은 요소

- ◆ 이 여러 블록 엘리먼트에 걸쳐있음
- ◆ <a> 가 닫히지 않음

브라우저별
결과는?



HTML 유효성검사



Opera Dragonfly ❖ B 엘리먼트의 자식 엘리먼트에 속함

```
<p>
  <strong>
    This text should be bold</p>
  <p>Should this text be bold? How does the HTML look when rendered?</p>
  <p>
    <a href="#">
      This text should be a link</p>
    <p>Should this text be a link? How does the HTML look when rendered?</p>
```

```
#text \r\n\r\n\t
-P
  -B
    #text This text should be bold
    #text \r\n\t
    -P
      #text Should this text be bold? How does the DOM look?
      #text \r\n\r\n\t
    -P
      -A href="#"
        #text This text should be a link
        #text \r\n\t
        -P
          #text Should this text be a link? How does the DOM look?
          #text \r\n\r\n\r\n
```

HTML 유효성검사



Mozilla Firebug

보이지 않는 `<b />` 엘리먼트가 추가됨

```
<p><strong>This text should be bold</strong></p>
<p><strong>Should this text be bold? How does the HTML look when rendered?</strong></p>
<strong></strong>
<p><a href="#">This text should be a link</a></p>
<p><a href="#">Should this text be a link? How does the HTML look when rendered?</a></p>
```

```
Console HTML CSS Script DOM Net
- <p>
  <b>This text should be bold</b>
  </p>
  <b/>
- <p>
  <b>Should this text be bold? How does the DOM look?</b>
  </p>
  <b/>
- <p>
  <a href="#">This text should be a link</a>
  </p>
  <a href="#" />
- <p>
  <a href="#">Should this text be a link? How does the DOM look?</a>
  </p>
```

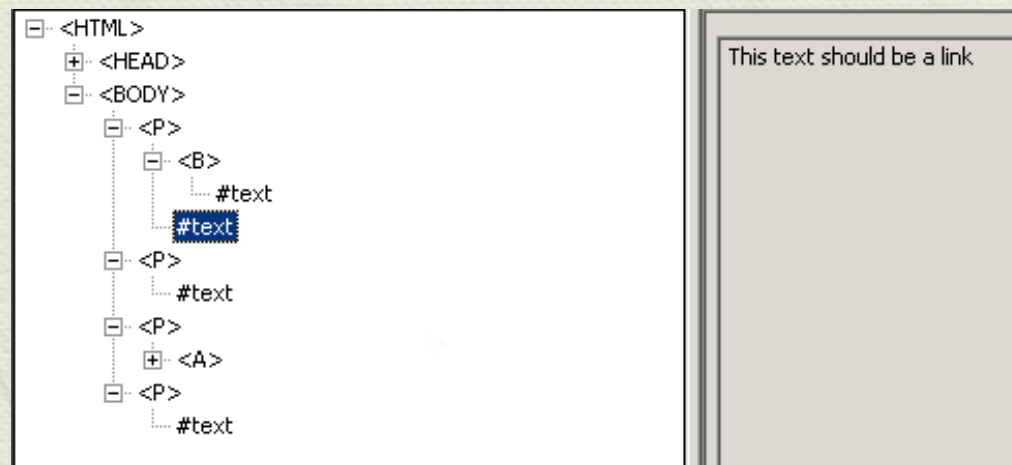
HTML 유효성검사



Internet Explorer ❖ This text should be a link 구문이 를 따라감

```
<p><strong>This text should be bold</strong>This text should be a link</p>
<p>Should this text be bold? How does the HTML look when rendered?</p>
```

```
<p><a href="#"></a></p>
<p>Should this text be a link? How does the HTML look when rendered?</p>
```



요약파싱기법으로 동적으로 생성되는 HTML의 유효성 검사하기

- ◆ 요약파싱기법
- ◆ HTML 유효성검사
- ◆ **진행과정**
- ◆ 더 해야할 것

기존 연구

◆ 동적으로 생성되는 HTML(혹은 XML)을 검사한 연구

- ◆ Y. Minamide, **Static approximation of dynamically generated web pages**, *WWW 2005*
 - ◆ 분석결과가 context-free grammar에 속하는지가 결정불가, 중첩되는 구조에 depth를 제한
- ◆ Y. Minamide and A. Tozawa, **XML validation for context-free grammars**, *APLAS 2006*
- ◆ C. Kirkegaard and A. Møller, **Static analysis for Java Servlets and JSP**, *SAS 2006*
 - ◆ Balanced grammar 사용. tag omission, inclusion, exclusion에 제약
- ◆ T. Nishiyama and Y. Minamide, **A translation from the HTML DTD into a regular hedge grammar**, *CIAA 2008*
 - ◆ Regular hedge grammar 사용해서 결정불가 문제해결, tag 관련 문제 제약

기존 연구

◆ 동적으로 생성되는 HTML(혹은 XML)을 검사한 연구

- ◆ K.-G. Doh, H. Kim, and D.A. Schmidt, **Abstract parsing: static analysis of dynamically generated string output using LR-parsing technology**, *SAS 2009*
 - ◆ 요약파싱기법, tag 관련 문제 고려하지 않음(HTML이 아닌 XHTML)
- ◆ A. Møller and M. Schwarz, **HTML validation of context-free languages**, *FOSSACS 2011*
 - ◆ SGML DTD 문법을 기준으로 파싱하는 방법, tag 관련 문제 해결

진행 과정

- ◆ JSP(+Java Servlet)가 생성하는 HTML 문서의 유효성 검사
 - ◆ 분석대상언어: JSP, 파서는 산학협력업체에서 제공
 - ◆ 참조문법: HTML 4.01 Transitional DTD를 LALR(1)로 변환
 - ◆ 의미분석 추가

진행 과정

- ◆ HTML 4.01 Transitional DTD를 LALR(1)로 변환하면서 문제

- ◆ LALR(1)로 표현하면서 모호함 발생: $LALR(1) \subseteq SGML$

HTML $\rightarrow$ HEAD BODY

HEAD $\rightarrow$ SCRIPT | ...

BODY $\rightarrow$ SCRIPT | ...

* HEAD와 BODY의 start, end tag가 모두 생략 가능

- ◆ 문법을 모호함 없이 느슨하게 작성하고 의미분석에서 엄격하게 검사

진행 과정

- ◆ 의미분석 추가

- ◆ 기존 유효성 검사기가 분석을 못하는 부분도 검사

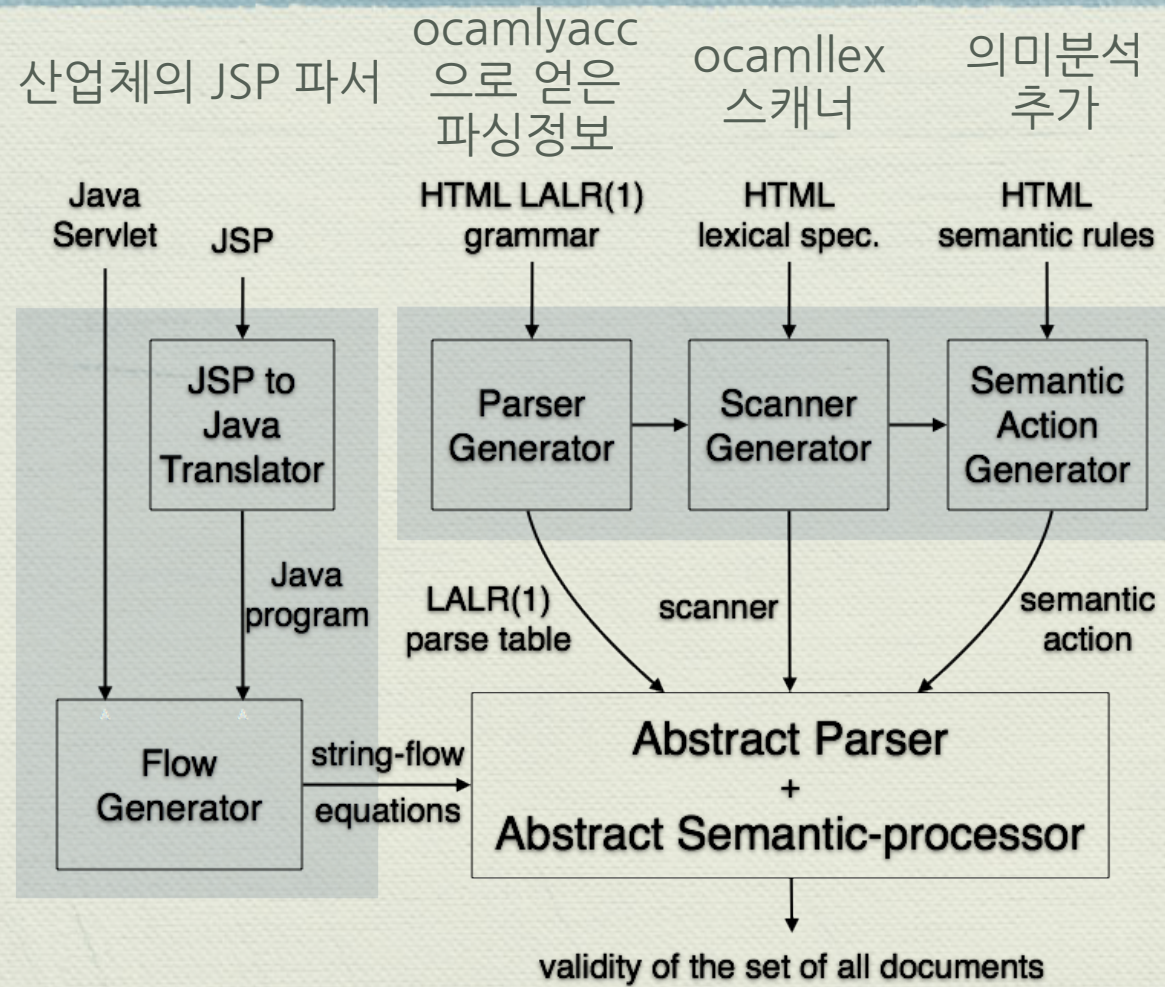
- ◆ 예) ID, IDREF 속성

- ◆ ID의 값은 유일해야 한다

- ◆ IDREF는 선언된 ID의 값이어야 한다

- ◆ SGML 문법을 LALR(1)로 표현하면서 모호한 부분을 해결

구조



요약파싱기법으로 동적으로 생성되는 HTML의 유효성 검사하기

- ◆ 요약파싱기법
- ◆ HTML 유효성검사
- ◆ 진행과정
- ◆ 더 해야할 것

더 해야할 것

- ◆ 논문을 위해서-
 - ◆ 좋은 예제 찾기 (의미분석이 의미가 있는 예제 찾기)
- ◆ 도구의 완성도를 위해서-
 - ◆ 파싱정보와 의미분석을 ocaml yacc 그대로 사용
- ◆ 연구를 위해서-
 - ◆ 요약파싱기법의 도메인:파싱그래프에 대한 정의

감사합니다