구조문서 환경에서 Annotation의 앵커링 기법

손원성°, 김재경, 최윤철, 임순범†
연세대학교 컴퓨터과학과
†숙명여자대학교 멀티미디어학과

Annotation Anchoring Methods in Structured Document Environments

Won-Sung Sohn, Jae-Kyung Kim, Yoon-Chul Choy, Soon-Bum Lim†,
Dept. of Computer Science, Yonsei University
†Dept. of Multimedia Science, Sookmyung Women's University
E-mail : {sohnws, ki187cm, ycchoy}@rainbow.yonsei.ac.kr, sblim@sookmyung.ac.kr

요 약

전자문서 환경에서의 annotation은 그 특성상 원본문서의 내용이 변경될 경우 annotation 의 대상인 앵커를 더 이상 참조할 수 없게 된다. 따라서 annotaton 시스템에서는 반드시 원본문서 변경에 대한 앵커링 기능을 필요로 한다. 그러나 기존 연구에서는 앵커 텍스트의 변경을 고려하지 않거나 일반 텍스트 문서만을 대상으로 한다. 본 논문에서는 XML과 같은 구조문서 환경에서의 annotation 앵커링 기법을 제안한다. 제안된 기법에서는 XML 환경에서 앵커 텍스트 및 path 정보에 대한 단계별 앵커링 과정을 수행한다. 또한 본 논문에서는 제안된 기법에 근거한 사용자 인터페이스를 제공한다. 그 결과 제안된 기법 및 시스템에서는 구조문서 환경에서 기존 연구 보다 심도 있는 앵커링을 보장하며 동시에 IETM, cyber-class, eLearing, semantic web 등의 다양한 분야에 효과적으로 적용 가능하다.

1. 서론

일반적으로 전자문서 환경의 annotation시스템에서는 annotation을 인라인(inline) 및 외부링크 형태로 생성하며, 그 결과 생성된 annotation 은 원본문서와는 별도로 시스템의 내부 및 외부에 저장된다. 이러한 특성상 annotation의 대상이 되는 원본문서의 내용이 삭제되거나 변경될 경우, 생성된 annotation은 더 이상 참조할 대상을 잃기 때문에 활용 가치 없는 고아가 된다. 따라서 이러한 문제는 반드시 해결되어야 하며 실제로도 annotation 시스템에서 사용자들이 가장 중요하게 간주하는 기능이기도 하다.

한편 대부분의 annotation 앵커링 기법에서는 unique id[3], substring[4], surrounding text[2] 등과 같은 컨텍스를 기본적으로 고려한다. 그러나 기존 연구에서는 anchor text의 변경을 고려하지 않거나[4], 일반 텍스트 문서만을 대상으로 한다[1]. 한편 annotation이 가장 빈번히 사용되는 cyber-class, eLearning, eBook 등의 환경에서는 대부분의 경우 XML을 원본 문서로 사용하므로, 구조 문서 환경에서의 robust한 앵커링 기법이 절실히 요구된다. 그러나 구조 문서환경을 고려하는 기존 연구에서는 생성된 annotation 및 원본 문서간의 간단한 path 비교만을 수행한다[2]. 본 논문에서는 XML과 같은 구조문서 환경에서의 annotation robust 앵커링 기법을 제안한다. 제안된 기법에서는 XML 환경에서 anchor text 및 path 정보에 대한 단계별 재 위치지정 과정을 수행한다. 먼저 원본문서 순회를 통한 생성된 annotation 정보와 문서 구조간의 일치 여부를 확인하여, 대상 문서의 갱신 정도를 판별한다. 다음 갱신 정도에 따라 각기 적합한 노드간의 대응관계를 생성하며, 이때에 발생하는 대응관계에 따라 단계별로 후보 앵커들을 생성한다.

또한 본 논문에서는 제안된 기법에 근거한 사용자 인터페이스를 제공한다. 본 인터페이스에서는 제안된 기법을 통하여 선택된 anchor 위치 정보를 제공하며 동시에 위 과정에서 제외된 후보 앵커들을 효과적으로 선택하기 위한 사용자와의 인터랙션 기능을 제공한다.

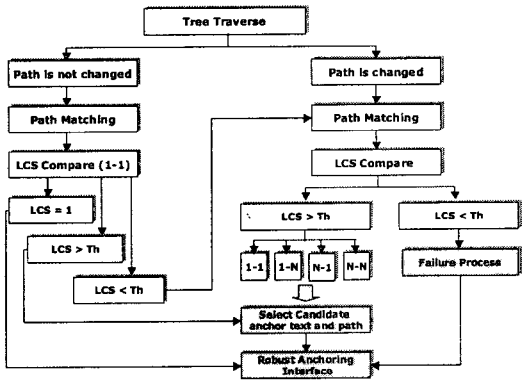
그 결과 제안된 기법 및 시스템에서는 구조문서 환경에서 기존 연구 보다 심도 있는 앵커링을 보장하며 동시에 IETM cyber-class, eLearing, semantic web 등의 다양한 분야에 효과적으로 적용 가능하다.

2. Annotaton의 앵커링 기법

본 논문에서는 XML과 같은 구조문서 환경에서의 annotation robust 앵커링 기법을 제안한다. 제안된 기법에서

는 먼저 원본문서 순회를 통하여 생성된 annotation 정보와 문서 구조간의 일치 여부를 확인하고, 대상 문서의 갱신 정도를 판별한다. 또한 갱신 정도에 따라 각기 적합한 노드간의 대응관계를 생성하며, 이때에 발생하는 대응관계의 단계에 따른 후보 앵커들을 생성한다.

이러한 제안기법의 전체 구성은 다음 그림 1과 같으며 각 단계별 상세설명은 계속하여 설명하도록 한다.



[그림 1] 제안기법의 전체 과정

2.1 패스 정보가 동일한 앵커 노드간의 Annotation 앵커링

제안기법에서는 그림 1 과 같이 원본문서의 구조와 annotation 간의 갱신여부에 따라 각기 다른 과정을 수행한다. 만일 annotation 패스정보와 원본 문서간의 패스정보가 동일하다면 일단 구조정보 보다는 anchor text 간의 갱신 여부를 판별하여야 한다.

한편 본 연구에서는 annotation 의 anchor text 와 원본문서의 텍스트 노드간의 갱신 여부를 판별하기 위하여 문자열의 유사도 비교에 사용되는 LCS(Longest common subsequence)에 기반한 LCS 의 비율인 식(1)을 이용한다.

$$lcsr(x,y) = \frac{2\times|lcs(x,y)|}{|x|+|y|}, \qquad (1)$$

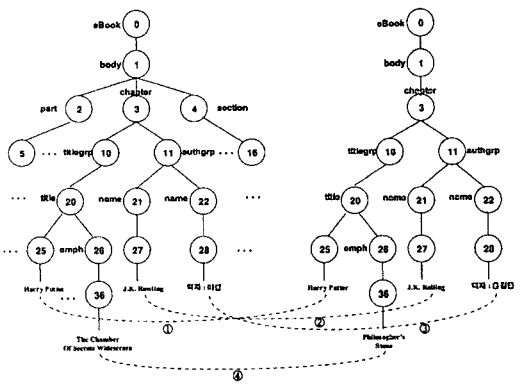
본 논문에서는 위와 같이 annotation 패스와 원본문서간의 구조가 갱신되지 않은 경우에는 다음 앵커링 기준 1 및 2 를 적용하도록 하며 자세한 내용은 다음과 같다.

앵커링 기준 1 : Annotation 정보인 T1 과, annotation 의 대상 문서인 T2 에 대하여 각각 annotation 의 앵커 텍스트 노드 T1 = $\{x_i\}$, $1 \le i \le s$, 와 원본문서의 텍스트 노드 T2 = $\{y_i\}$, $1 \le i \le \rightarrow$, 가 존재한다고 하자. 또한 각각의 텍스트 노드들은 문자열 x_i= {T1*strin*}, $1 \le n \le \lambda$ 및 y_i= { T2*strin* }, $1 \le n \le \downarrow$을 포함하고 각각의 문자열들은 캐릭터 T1str_{ik}= $\{\alpha_{ik}\}$, $1 \le n \le \kappa$, 및 T2str_{ik}= $\{\beta_{ik}\}$, $1 \le n \le °$, 를 포함한다. 이때에 annotation 정보 T1 의 모든 노드가 레이블 및 순서가 동일한 형태로 T2 에 존재하고, 부모노드의 레이블이 동일한 각각의 T1 의 앵커 텍스트 노드 $x_{i1},\ldots,x_{ip}$ 와 T2 의 텍스트 노드 $y_{i1},\ldots,y_{ir}$ 가 존재한다면, 대응관계 $[x_{i1}, y_{i1}],\ldots,[x_{ip}, y_{ir}]$를 생성한다.

탐지 기준 1 을 만족한다면 원본 문서의 패스는 변경되지 않았다고 가정할 수 있기 때문에 각 anchor text 간의 일-대-일 대응관계 생성이 가능하다. 예를 들어 다음 그림 2 와 같이 원본문서의 구조가 변경되지 않은 경우에는 원본문서와 annotation 패스간의 일-대-일 대응관계인 대응관계 [25, 25], ①, 및 [27, 27], ② , [28, 28], ③, 그리고 [36, 36], ④ 의 생성이 가능하다.

앵커링 기준 1 에 의해서 생성된 대응관계가 생성되었다면, 각 anchor text 간의 유사도를 측정하고 그 결과에 따라 annotation 앵커링을 설정한다. 이에 대한 내용은 다음 기준 2 에서 설명하고 있다.

앵커링 기준 2 : 기준 1 에 의하여 생성된 대응 관계간 노드들의 문자열 유사도가 1 인 경우는 원본 문서의 텍스트 노드 $y_{i1},\ldots,y_{in}$ 를 앵커 영역으로 지정한며, 만일 유사도가 1 이하이면서 일정 임계값 이상일 경우, 텍스트 노드 $y_{i1},\ldots,y_{in}$ 를 후보 앵커 영역으로 지정하여 별도의 위치 선정 과정을 수행한다. 한편 일정 임계값 이하인 경우는 삭제되거나 경로가 변경되었다고 가정하여 앵커링 기준 3~6 들을 적용한다. 만일 기준 3~6 을 적용하여 조건을 만족하는 T2 의 텍스트 노드가 존재하지 않는다면 후보 앵커 지점을 T2 의 텍스트 노드 y_{in} 로 지정한다.



[그림 2] 앵커링 기준 1 및 2 의 적용 결과

탐지 기준 1, 2 를 적용한 결과 패스가 변경되지 않고, 앵커 텍스트가 동일 혹은 갱신되었을 경우 기존 영역을 그대로 사용하거나, 새로운 후보 앵커 영역을 추출할 수 있다.

[1] |lcs(x,y)|, |x|, 그리고 |y|는 각각 문자열 x와 y간의 LCS의 길이, x의 길이, 그리고 y의 길이를 나타낸다.

예를 들어 그림 2 의 대응관계 [25, 25], ①, 은 유사도 1 의 관계이기 때문에 바로 annotation anchor 로 지정 가능하며, 대응관계 [27, 27], ② ,와 [28, 28], ③, 은 유사도가 1 이하이면서 일정 임계값 이상을 포함하므로 앵커 노드 [27, 27], 과 [28, 28], 을 후보 앵커영역으로 지정한다. 한편 대응관계 [36, 36], ④, 은 유사도가 임계값의 기준을 벗어나므로 다음에서 설명할 새로운 앵커 기준을 적용하도록 한다. 만일 대응관계 [36, 36], ④, 가 새로운 앵커링 기준을 만족하지 않는다면 이는 anchor text 가 완전히 수정된 경우이므로, <emph> 엘리먼트의 텍스트 노드를 후보 앵커 지점으로 선택한다. Annotation 앵커링 기준으로 선택된 후보 앵커들은 제안 인터페이스의 사용자 인터랙션을 통하여 최종 annotation anchor 로 선택된다.

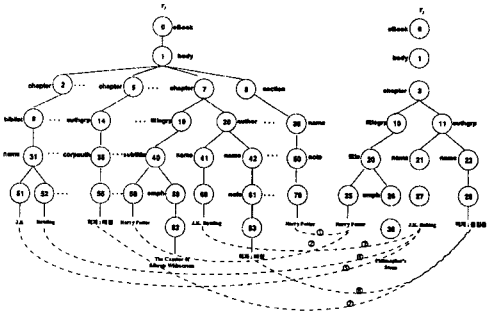
2.2 패스 정보가 상이한 앵커 노드간의 Annotation 앵커링

원본문서와 annotation 간의 패스 및 앵커 텍스트의 변화가 발생하였을 경우는 2.1 절의 기준과 같이 anchor text 간의 비교 뿐만이 아니라 구조간의 유사성 등도 고려할 수 있어야 한다. 또한 이러한 정보에 의하여 추출한 후보 앵커들은 annotation 특성 상 앵커 텍스트 및 패스에 대한 일-대-일, 일-대-다, 다-대-다, 다-대-일과 같은 복수의 후보들을 포함한다. 따라서 본 연구에서는 이러한 복수의 후보들에 대한 적절한 선택을 위하여 단계별 경로 매칭 및, 복수 후보간의 병합, 링크 등의 기법들을 적용한다. 상세한 설명은 다음과 같다.

앵커링 기준 3 : DOM 트리 순회를 통하여 Annotation 정보인 T_1 의 패스 정보가, annotation 의 대상 문서인 T_2 의 노드 레이블 및 형제 순서 등과 일치하지 않는 경우가 발생한다고 하자. 이때에 annotation 앵커 텍스트 노드인 xio,..., xip 와 T2 의 모든 텍스트 노드간의 문자열 유사도를 추출하여 일정 임계값 이상이 되는 T2 의 텍스트 노드를 선택하여 대응관계 [xio, yiq],...,[xio, yir],...,[xip, yiu],...,[xip, yiv] 등을 생성한다.

원본문서에서 구조 및 텍스트 노드의 내용이 갱신되었다면 먼저 앵커링 기준 3 을 이용하여 문자열 유사도가 일정 이상되는 대응관계들을 생성한다. 다음 그림 3 처럼 노드 25 에 대해서는 대응관계 [25, 70], ①, 및 [25, 58], ②를 생성하며, 같은 기준에 근거하여 대응관계 ③, ④, ⑤, ⑥, ⑦ 을 생성할 수 있다.

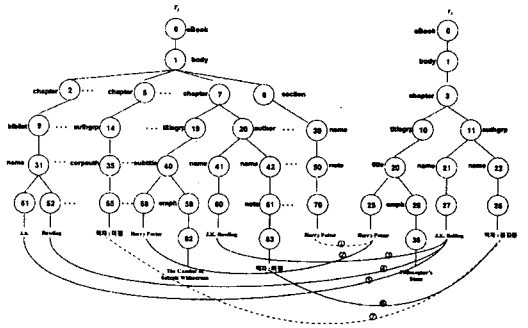
한편 기준 3 에 의하여 생성된 대응관계들은 그림의 경우처럼 대부분 복수의 앵커 노드 대 복수의 텍스트 노드들로 대응된다. 제안기법에서는 이러한 경우 단순히 문자열 유사도에 근거하여 하나의 대응관계를 선택하기 보다는, 노드들의 구조적 연관성을 추출하여 새로운 대응관계를 생성하여, 그 결과를 보다 의미있는 앵커로 간주한다. 이러한 내용은 다음 기준 4 와 같다.



[그림 3] 문자열 유사도에 의한 앵커 후보 생성

앵커링 기준 4 : Annotation 의 앵커 노드에 대한 원본문서의 텍스트 노드문자열 유사도가 일정 이상이 되는 대응관계들이 복수로 존재한다면, 대응하는 각각의 텍스트 노드의 경로[2] 간 레이블들 유사도 []를 비교하여 일정값 이상의 레이블 유사도를 포함하는 노드들의 대응관계[xio, yiq],...,[xio, yir],...,[xip, yiu],...,[xip, yiv] 를 생성한다.

기준 4 에 의하여 문자열 유사도에 근거한 복수의 대응계들 중 레이블 유사도가 일정 이상인 대응 관계들은 새로운 후보 앵커로 지정된다. 예를 들어 다음 그림 4 에서 검은색 라인의 대응관계 [25, 58], ②, [27, 60], ③, [27, 52], ④, [27, 51], ⑥, 그리고 [28, 83], ⑦ 은 경로간 레이블 유사도가 일정 이상이 되는 경우들이며, 이들을 새로운 대응관계로 지정한다. 따라서 기준 4 의 대응관계들은 기준 3 에 의하여 생성된 대응관계 보다 높은 우선 순위의 후보 앵커로 간주할 수 있으며 이러한 정보는 차후 앵커링 interface 에서 사용자 인터랙션의 기준으로 사용한다.



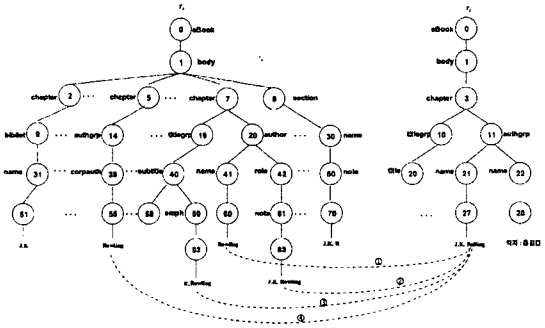
[그림 4] 레이블 유사도에 의한 앵커 후보 생성

한편 제안기법에서는 앵커링 interface 에서 보다 다양한 recommend 노드를 제공하기 위하여 경로간 레이블 유사도의 임계값을 만족하지는 않지만 문자열 유사도가 어느 정도 유사한 노드들에 대해서도 후보 앵커로 간주한다. 이를 위해서는 다음 기준 5 를 적용한다.

[2] 경로(x)란 노드 x의 부모 노드로부터 뿌리노드까지의 노드의 순차적인 집합을 의미한다.

앵커링 기준 5 : 만일 T2 의 텍스트 노드에서 기준 3 을 만족하지만 기준 4 를 만족하지 않는 대응관계들이 존재한다면 이러한 대응관계들의 부모노드가 서로 동일한 노드들을 우선적으로 후보 앵커로 간주하고 나머지 대응관계들에 대해서는 유사도 값을 기준으로 후보 앵커를 지정한다.

다음 그림 5 에서 T2 의 앵커 노드 27 에 대한 대응관계 중 문자열간 유사도가 일정이상이면서 경로간 레이블 유사도가 일정 이하인 대응관계는 [27, 55], ① , [27, 60], ②, [27, 82], ③, 그리고 [27, 83], ④와 같다. 이러한 복수의 대응관계중에서 먼저 부모노드가 일치하는 대응관계 ①을 가장 우선적으로 후보 앵커로 선택하며, 다음으로 유사도를 기준으로 대응관계 ②, ③, ④를 차례로 선택한다.

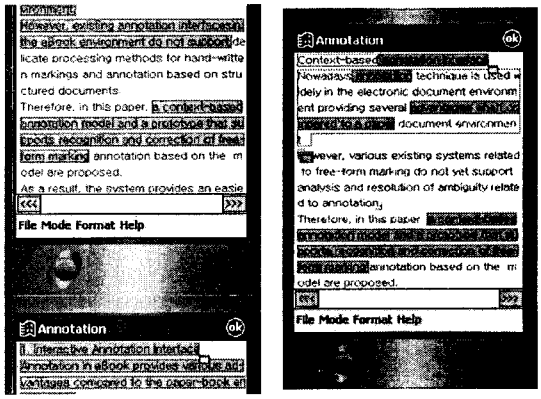


[그림 5] 기준 5 를 적용한 결과

3. 구현결과

본 절에서는 제안 기법 및 인터페이스를 포함한 annotation 시스템을 설명한다. 본 시스템은 XML 기반의 EBKS 전자책표준을 사용하였으며, Window XP 및 Window CE 환경에서 구동된다. 본 내용에서는 Window CE 환경에서 실행되는 annotation 시스템을 설명한다.

먼저 다음 그림 6 (A)에서는 제안 시스템의 annotation 브라우저에서 입력된 annotation 정보들을 나타내고 있으며 총 4개의 하이라이팅과 1개의 노트를 삽입하였다. 그림 6 (B)에서는 변경된 그림(A)의 대상문서에 대하여 제안 기법 및 인터페이스를 적용하여 robust 앵커링을 제공하고 있는 내용을 보여주고 있다. 그림 6 (B)에서는 먼저 그림 (A)의 annotation에 대한 primitive 앵커링을 박스(rectangle) 형태의 annotation으로 출력하며 특히 문자열 유사도에 의하여 추출된 공통된 anchor text를 하이라이트로 포함한다. 이는 사용자에게 앵커링 선택의 기준을 제공하기 위함이다. 또한 그림 6 (A)의 세번째 annotation의 경우는 그림 6 (B)에서 앵커의 구조는 동일하지만 anchor text의 내용이 심하게 변경되었기 때문에 제안 인터페이스에 근거하여 앵커노드의 엘리먼트 영역을 primitive anchor로 출력하고 별도의 아이콘을 통하여 기존 annotation의 내용을 제공한다.



(A) (B)

[그림 6] 원본문서(A) 변경에 대한 앵커링 결과(B)

4. 결론

본 논문에서는 구조문서 환경에서 robust한 앵커링 기법을 제시하였다. 제안된 기법에서는 XML 환경에서 텍스트 및 path 정보에 대한 단계별 재 위치지정 과정을 수행한다. 먼저 원본문서 순회를 통하여 대상 문서의 갱신 정도를 판별하고 갱신 정도에 따라 각기 적합한 노드간의 대응관계를 생성하며, 대응단계별로 후보 앵커들을 생성한다.

그 결과 본 논문의 제안된 기법 및 시스템을 통하여 구조문서 환경에서 기존 연구 보다 심도 있는 앵커링을 보장하며 동시에 IETM, cyber-class, eLearing, semantic web 등의 다양한 분야에 효과적으로 적용 가능하다.

[참고문헌]

[1] Brush, A., Bargeron, D., Gupta, A., and Cadiz, J. Robust Annotation Positioning in Digital Documents. Proc. CHI 2001, 285-292.

[2] Phelps, T., and Wilensky R. Robust Intra-document Locations, Proc. of the 9th WWW Conference, http://www9.org-/w9cdrom/312/312.html

[3] Rizk, A. & Sauter, L. Multicard: An Open Hypermedia System. In: D. Lucarella, J. Nanard, M. Nanard, P.Paolini. eds. The Proceedings of the ACM Conference on Hypertext, ECHT '92 Milano, pp 181-190. ACM. 1992

[4] Röscheisen, M., Mogensen, C., and Winograd, T. Shared Web Annotations as a Platform for Third-Party Value-Added, Information Providers: Architecture, Protocols, and Usage Examples, Technical Report CSDTR/DLTR (1997), Stanford University.