제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

개선된 Spreading Activation을 이용한 객체지향 컴포넌트의 검색

김귀정*
*건양대학교 IT학부
e-mail:gjkim@konyang.ac.kr

Retrieval of Object-Oriented Component using Enhanced Spreading Activation

Gui-Jug Kim*
*Division of Information Technology, KonYang University

요 약

본 연구는 객체지향 컴포넌트 검색을 위해서 개선된 Spreading Activation 방법을 이용하여 다중 패싯 분류된 컴포넌트를 효율적으로 검색할 수 있는 방법을 제안하였다. 객체지향 코드 기반의 관계정의를 위해 특성과 컨텍스트 간에 연관관계를 설정하고, 컨텍스트의 자동 추출을 위한 Spreading Activation 방법의 초기 활성값을 정의하였다. 쿼리에 대해 자동 검색된 컨텍스트에 의해 후보컴포넌트가 선정되고, 쿼리와 컴포넌트 간의 신뢰도가 계산됨으로써 컴포넌트가 검색될 수 있도록 하였다. 본 연구는 다중 패싯 분류된 객체지향 컴포넌트의 검색에 효율적이며, 사용자 수작업의 부담을 최대한 감소시켜 컴포넌트의 재사용성을 높일 수 있도록 하였다.

1. 서론

소프트웨어 생산성과 유지보수 비용을 줄일 수 있는 기법으로 다양한 컴포넌트 기반의 개발 방법론이 제안되고 있으며, 현재 한국컴포넌트컨소시엄(KCSC)등에서 상호호환을 위한 컴포넌트 개발에 관한 연구가 진행되고 있다. 이와 같은 많은 컴포넌트가 개발되었을 경우, 사용자가 원하는 컴포넌트 식별방법은 분석자의 경험에 의존하는 경우가 많다. 특히 다중 카테고리에 분류된 객체지향 컴포넌트의 경우에는 대부분 개발자의 경험에 의해 검색이 이루어지고 있다. 이러한 검색방법은 주관적이고, 비효율적이며, 시스템의 일관성을 유지하기도 어렵다. 이에 통합환경에서의 인프라를 제공해야 하며 커스터마이징(Customizing)을 위한 정확하고 자동화된 컴포넌트의 검색 방법이 필요하다.

본 연구에서는 CBSE(Component-Based Soft ware Engineering) 즉, 신속한 시스템 구축, 비용절감을 위하여 재사용이 가능하도록 하는 객체지향 컴포넌트의 검색에 목적을 둔다. 컴포넌트는 컨텍스트에 의해 패싯 분류되고, 각 컨텍스트는 특성과 연관성을 가지게 된다. 이 연관성은 Spreading Activation 방법의 적용을 가능케 해주고, 이로 인해 컨텍스트의 자동 검색이 이루어질 수 있도록 하였다. 각 특성과 컨텍스트에 대한 초기 활성값을 정의하였으며, 쿼리에 대한 컴포넌트와의 신뢰도를 측정하여 최적의 컴포넌트를 검색할 수 있도록 하였다. 따라서 사용자가 원하는 컴포넌트를 쉽게 선택할 수 있도록 하고 또한 사용자 수작업의 부담을 최대한 감소시켜 컴포넌트의 재사용성을 높일 수 있도록 하였다.

2. 관련 연구

기존의 컴포넌트 검색 연구는 시그니쳐 일치 검색[1], 행위 샘플링에 의한 검색[2], 명세서 일치에 의한 검색[3] 등이 있다. 시그니쳐 일치 검색은 함수의 파라메터 타입이나 인터페이스와 같은 시그니쳐 정보를 이용하여 컴포넌트를 검색하는 방법이며, 행위 샘플링 검색은 인터페이스 명세서와 호환되는 인터페이스를 가진 저장소의 루틴을 실행하여 그 결과를 비교함으로써 검색하는 방법이다. 명세서 일치 방법은 컴포넌트 명세서를 비교하여 다른 컴포넌트와 대체될 수 있는지를 비교하여 결정한다.

SARM(Spreading Activation Retrieval Method)[4]은 컴포넌트와 질의어 사이에 질의어 기능을 포함하는 유사한 컴포넌트들을 검색하여 보다 더 정확하고 넓은 범위의 컴포넌트들을 찾을 수 있는 방법이다. SARM은 직접 인덱싱되지 않은 컴포넌트들까지 검색할 수 있는 효율적인 검색 방법이며 정보저장소에 컴포넌트들을 구축할 때 각 항목을 일일이 인덱싱하지 않아도 되기 때문에 많은 비용이 절감된다. 그러나 활성값을 이용한 반복 계산으로 유사도를 측정하였기 때문에 검색시간을 지연시키는 단점이 있다.

3. 개선된 Spreading Activation을 이용한 컨텍스트검색

3.1 Spreading Activation을 위한 관계정의

본 연구에서는 패싯분류의 개념을 사용하여 컴포넌트를 하나 이상의 컨텍스트(context)로 분류하였다[5]. 컨텍스트는 소스 코드로부터 추출된 특성(feature)으로 구성되는데, 특성은 메소드 이름에 해당하는 동사형태와 메소드의 첫 번째 인수에 해당하는 명사형태의 한 쌍으로 이루어져 있다. 그림 1은 컴포넌트의 구조를 나타낸다.

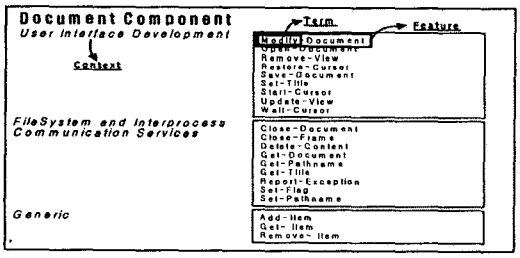


그림 1 컴포넌트 구조

컨텍스트에 의해 다중 패싯 분류된 컴포넌트의 검색을 위해서는 쿼리가 표현하고자 하는 컨텍스트가 어떤 것인가를 찾아야 하고, 또한 그 쿼리가 여러 개의 컨텍스트를 만족할 수 있음을 이해해야 한다. 이에 본 연구에서는 쿼리가 동사, 명사의 특성형태로 주워졌을 때, 이를 만족하는 컨텍스트를 찾기 위하여 Spreading Activation 알고리즘을 이용하였다[6]. 이를 위하여 특성들과 컨텍스트들의 초기 활성값을 설정하였다.

각 특성의 초기 활성값은 컴포넌트와 특성간의 가중치에 의해 결정된다. 이를 특성 가중치(Feature Weight)라 하고, 이는 컴포넌트의 행동에 가장 밀접한 특성을 강조하는 역할을 한다. 특성 가중치에 대한 수식은 식(1)에 나타나 있다. 각 특성의 초기 활성값은 한 특성에 대해서 계산된 특성 가중치의 평균으로 설정한다.

$$FW_{i,k}=\frac{ff_{i,k}\times\log(\frac{N}{n_k})}{\sqrt{\sum_{z=1}^{F}(ff_{i,z}\times\log(\frac{N}{n_z}))^2}} \qquad \text{식(1)}$$

$FW_{i,k}$: i번째 컴포넌트의 k번째 *feature*의 *weight*
N : 전체 컴포넌트의 수
$ff_{i,z}$: *feature frequency*
n_k : k번째 *feature*가 나타나는 컴포넌트의 수
F : *repository*에 있는 서로 다른 *feature* 수

컨텍스트의 초기 활성값을 설정하기 위해서 특성과 컨텍스트 간의 연관성를 이용한 특성-컨텍스트관계값(Feature-Context Value)을 정의한다. 특성-컨텍스트관계값에 대한 수식은 식(2)에 나타나 있다. 이는 각 컨텍스트에 속한 특성의 수는 컴포넌트와 컨텍스트와의 관련성을 암시해 준다는 의미에 근거한다.

$$FCV_{l,j}=p_f(l)\frac{feature_{l,j}}{feature_l} \qquad \text{식(2)}$$

$FCV_{l,j}$: *Context* j와 *feature* l과의 관계값
$p_f(l)$: *Context* j에서의 l 번째 *feature*의 발생백분율
$feature_{l,j}$: *Context* j에서 l 번째 *feature*의 발생횟수
$feature_l$: 모든 *Context*에서 l 번째 *feature*의 전체발생횟수

특성-컨텍스트관계값을 바탕으로 특성과 특성 사이의 일치도를 계산할 수 있다. 먼저, 하나의 컨텍스트에 대하여 특성 A와 특성 B 사이의 일치값을 식(3)에 따라 계산한다. 그후, 모든 컨텍스트에 대해서 계산한 후 그 값을 합산한다(식(4)). 이를 특성-특성 일값치값(Feature-Feature Value)이라 정의하고, 이 값은 검색 시 쿼리의 각 특성과 후보 컴포넌트의 특성 사이의 동치관계 계산식에 이용한다.

$$mj_{ab}=\frac{1}{1+|a1-b1|}\ ,\quad (a1\neq 0,\ b1\neq 0)$$
$$mj_{ab}=0\ ,\quad (a1=0,\ OR\ b1=0)\ \text{또는}\ (a1=b1=0) \qquad \text{식(3)}$$

$$M_{AB}=\sum_{j=1}^{c}mj_{ab} \qquad \text{식(4)}$$

3.2 컨텍스트 자동검색

개선된 SARM(Enhanced Spreading Activation Retrieval Method)은 순환과정이 일정 수준 반복된 후 기준에 미치지 못하는 컴포넌트들의 연결정보를 제거하여 연산에서 제외시킴으로써 검색의 확장범위를 줄여 보다 관계가 깊은 후보컴포넌트들만을 검색한다. 이를 컨텍스트의 자동검색에 적용하기 위하여 특성과 컨텍스트 사이에 부여된 연관성을 이용한다. 각 특징의 초기 활성값은 한 특성에 대해서 계산된 특성 가중치(FW)의 평균값이고, 컨텍스트의 초기

활성값은 특성과 컨텍스트 간의 연관성를 이용하여 계산된 특성-컨텍스트관계값(FCV)이다. 이를 바탕으로 동사-명사의 특성 형태로 주어진 쿼리에 대해서 컨텍스트를 검색하는 과정이 그림 2에 나타나 있다. 특성 A의 초기 활성값은 0.8이고, 컨텍스트 W의 초기 활성값은 0.5이다. 쿼리로 특성 「A」를 입력하면 3개의 컴포넌트가 검색된다. 여기서 특성 「A」는 컨텍스트 「W」와 「X」에 직접 연결되어 있지만 컨텍스트 「Y」와 「Z」에는 연결되어 있지 않다. 그러나 「A」→「X」→「D」→「Z」를 통하여 연결되고, 「「A」→「X」→「D」→「Y」를 통하여 2개의 컨텍스트(「Z」,「Y」)가 연결됨을 알 수 있다. 하지만 「Y」는 검색과정에서 적게 참조되므로 연결이 제거된다. 이처럼 각 특성와 컨텍스트는 서로 연결되어 있는 노드를 참조해 가면서 활성값을 계산하게 된다. 순환이 반복될수록 활성값은 안정되며 참조회수가 기준에 미달되는 부분은 자동으로 제거되어 계산과정이 종료된다.

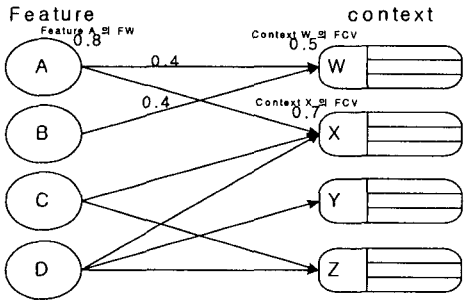


그림 2. 쿼리에 대한 컨텍스트 검색 과정

4. 객체지향 컴포넌트 검색

4.1 검색 시스템 구조

본 연구에서 제안한 검색은 객체지향 컴포넌트의 재사용과 역공학을 위한 코드 이해도를 높이는데 그 목적이 있다. 이는 사용자가 익숙한 방법으로 쿼리를 제공하고 최소한 수작업을 줄여 자동으로 컴포넌트를 검색함으로서 구현될 수 있다. 그림 3은 검색 과정을 보여준다. 특성 형태로 주워진 쿼리는 시소러스에서 계산되어진 특성가중치와 특성-컨텍스트관계값을 이용하여 쿼리와 연관된 컨텍스트를 자동 추출한다. 이때, 컨텍스트의 자동 추출을 위해서는 Spreading Activation 방법이 사용되어져 쿼리와 연관된 후보컨텍스트들이 추출된다. 그 후 쿼리셋에서 공통적으로 추출된 컨텍스트를 만족하는 컴포넌트를 찾는다. 찾아진 후보컴포넌트와 쿼리와의 신뢰도가 계산되고, 최종적으로 신뢰도의 우선순위로 컴포넌트가 검색된다.

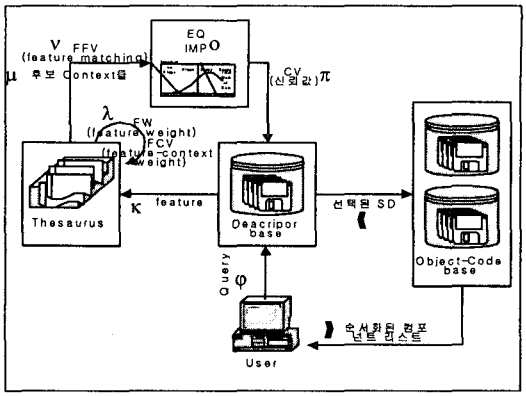
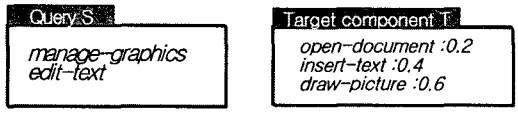


그림 3. 컴포넌트 검색 과정

4.2 신뢰도에 의한 컴포넌트 검색

컨텍스트 검색 결과, 한 쿼리셋에서 공통적으로 나타나는 컨텍스트를 모두 만족하는 후보컴포넌트에 대해서 쿼리셋과 컴포넌트들과의 신뢰도를 계산한다. 최종적인 신뢰도는 동치관계, 포함관계, 유사도를 계산함으로써 얻어진다. 다음의 예를 통해 쿼리셋과 컴퍼넌트간의 신뢰도 계산과정을 알아본다.



쿼리S의 특성가중치(FW)가 각각 0.7, 0.8이라 하자.

① 쿼리S와 대상컴포넌트의 각 특성에 대한 동치관계(Equivalence)를 계산한다. 동치관계는 두 특성이 얼마나 유사한가를 나타내주는 값이다. 이 값은 앞서 정의한 특성-특성일값치값(Feature-Feature Value)에 의해 구해질 수 있다.

$$Eq(S(u), T(v)) = FFV(S(u), T(v)) \qquad \text{식(5)}$$

식(5)에 의해서 얻어진 값이 표 1에 나타나 있다.

표 1. 동치관계

EQ	open-document	insert-text	draw-picture
Manage-graphic	0.2	0.3	0.2
Edit-text	0.3	0.3	0.3

② 특성가중치와 동치관계를 이용하여 포함관계(Implication)를 계산한다. 포함관계는 두 특성이 교환될 수 있는 정도를 나타낸다.

$$Imp(S(u), T(v)) = \min(1, \max FW(T(v)), FW(S(u)))[EQ(u, v)] \qquad \text{식(6)}$$

식(6)에 의해서 얻어진 포함관계 값은 표 2와 같다.

표 2. 포함관계

IMP	open-document	insert-text	draw-picture
Manage-graphic	0.14	0.21	0.14
Edit-text	0.24	0.24	0.24

③ 정규화된 가중치 벡터를 이용하여 유사도(Simmularity)를 계산한다.

$$SIM=(IMP^T * EQ) * normalized\ weight\ vector \quad 식(7)$$

$SIM=\{0.1,\ 0.135,\ 0.1\}*\{FW(T(v))/FWT(1)+FWT(2)+FWT(3)\}$
$= \{0.016,\ 0.043,\ 0.05\}$

④ 조절함수에 의해 최종적으로 신뢰도(Confidence Value)가 계산된다.

$$CV=10\sum_{i=1}^{t} D_i CV_i \quad 식(8)$$

$CV = 10*\{0.016,\ 0.043,\ 0.05\}*\{0.5,1,0.5\} = 0.76$

5. 성능 평가

본 연구는 다중 패싯 분류된 컴포넌트의 검색 효율을 높이기 위하여 특성을 기본으로 한 검색방법을 제안하였다. 이를 위하여 특성가중치와 특성-컨텍스트관계값을 이용한 Spreading Activation 방법을 적용하였다. 검색 시스템의 재현율과 정확성을 측정하기 위해서 쿼리에 대한 컴포넌트 검색 결과를 실험하였다. 쿼리는 임의로 30개를 선정하고 각각의 쿼리 대한 컴포넌트 검색 결과를 측정하였다. 그림 4와 그림 5는 이 결과에 대한 정확도와 재현율을 측정한 것이다. Spreading Activation 방법을 적용하지 않았을 때와, 제안한 방법을 비교하였다. 정확도 면에서도 뒤지지 않으면서 재현율이 크게 향상되었음을 알 수 있다.

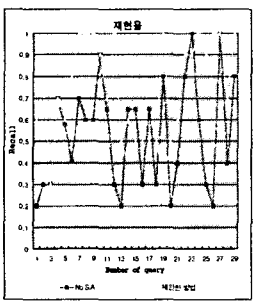


그림 4. 재현율

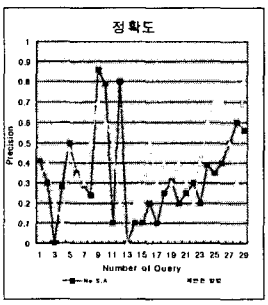


그림 5. 정확도

6. 결론

본 연구는 개선된 Spreading Activation 방법을 이용하여 객체지향 컴포넌트를 효율적으로 검색할 수 있는 방법을 제안하였다. 이를 위해 특성과 컨텍스트 간의 관계를 설정하였고, 이 연관관계는 컨텍스트의 자동 추출을 위한 Spreading Activation 방법의 초기 활성값으로 이용될 수 있도록 하였다. 본 연구는 기존의 검색 시스템과 비교하여 다중 패싯 분류된 컴포넌트의 검색에 효율적이며, 검색 시 컨텍스트를 직접 선택해야 하는 사용자 수작업의 부담을 최대한 감소시켜 컴포넌트의 재사용성을 높일 수 있었다.

앞으로의 연구 방향은, 컨텍스트의 상속 관계 표현과 프레임워크 라이브러리로 확장시키는 방법이 요구된다.

참고문헌

[1] A. M. Zaremski, J. M. Wing, "Signature Matching: A Tool for Using Software Libraries," ACM Transaction Software Engineering and Methodology, Vol. 4, No. 2, 1995.

[2] A. Podgurski, L. Pierce, "Retrieving Reusable Software by Sampling Behavior," ACM Transaction Software Engineering and Methodology, Vol. 2, No. 3, 1993.

[3] A. M. Zaremski, J. M. Wing, "Specification Matching of Software Components," In Proceedings of the third ACM SIGSOFT symposium on the foundations of software engineering, 1995.

[4] Scott Henninger, "Information Access Tools for Software Reuse," System Software, pp. 231-247, 1995.

[5] E. Damini, M.G.Fugini, C. Bellettini, "A Hierarchy-Aware Approach to Faceted Classification of Object-Oriented Components", The ACM Transaction on Software Engineering and Methodology, Vol.8, No.4, Oct. 1999, 425-472.

[6] 한정수, 송영재, "개선된 SARM을 이용한 객체지향 부품 재사용 시스템," 정보처리논문지, 제7권 제4호, pp. 1092-1102, 4. 2000.