Case 기반 컴포넌트 검색 시스템 설계

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

Case-Based Retrieval System Construction

Gui-Jug Kim*
*Dept. of Computer Engineering, KonYang University

요 약

본 연구는 소스 코드를 재사용하기 위한 Case 기반 검색에 있어서 효율적인 검색 시스템을 구축하는 방법을 제안하고자 한다. 소스 코드의 객체지향적인 특성을 만족하기 위하여 각 노드 간 객체지향 상속의 개념을 표현할 수 있도록 초기 관련값을 시소러스로 구축하고자 한다. 이때, 각 Case를 구성하는 클래스들을 상속관계에 따라 개념적으로 분류하였고, 시소러스 방법에 퍼지 논리를 적용하여 객체지향 시소러스를 생성하여 의미망을 구축한다. 또한, 의미망의 노드와 간선을 활성화시키고 활성값을 전파시키기 위해 사용되는 spreading activation 방법의 단점을 보완하여 spreading activation의 성능은 최대한 유지하면서 검색 속도를 향상시킬 수 있는 방법을 제안하고자 한다.

1. 서론

객체지향 언어와 같은 프로그래밍 언어는 다양한 종류의 재사용 컴포넌트를 사용한다. 이러한 라이브러리는 일반적으로 대단히 방대하다. 예를 들어, 자바의 경우 표준 클래스 라이브러리에 약 3000개의 클래스와 인터페이스가 포함되어 있다. 그러므로 컴포넌트 재사용을 효율적으로 하기 위해서는 방대한 컴포넌트 관리와 검색이 필요하다. 또한 컴포넌트를 검색하는 방법에 따라 사용자의 요구사항을 얼마나 반영할 수 있는가와 라이브러리 확장성 등이 결정되기 때문에 검색 방법은 매우 중요하다. 특히 소스 코드를 재사용하기 위한 Case 기반 검색에 있어서 의미망의 효율적인 구성은 무엇보다 중요하다[1].

이에 본 연구에서는 의미망의 노드와 간선을 활성화시키고 활성값을 전파시키기 위해 사용되는 spreading activation 방법의 단점을 보완하여 spreading activation의 성능은 최대한 유지하면서 검색 속도를 향상시킬 수 있는 검색 시스템을 설계하고자 한다. 또한, 각 노드 간 객체지향 상속의 개념을 표현할 수 있도록 초기 관련값을 시소러스로 구축하고자 한다. 이때, 각 Case를 구성하는 클래스들을 상속관계에 따라 개념적으로 분류하였고, 시소러스 방법에 퍼지 논리를 적용하여 객체지향 시소러스를 구축한다.

2. Case 기반 재사용

많은 CBR 시스템에서 라이브러리 내에 있는 각 클래스를 케이스(Case)로 취급하고 있으며, 또한 이러한 Case를 재사용 단위인 컴포넌트화 하고 있다[2]. Case는 각 컴포넌트의 소스 코드를 가지고 있으며, 그 컴포넌트를 표현할 수 있는 속성이나 특징을 포함할 수 있다. 또한 컴포넌트의 행위적 특성을 텍스트로 포함할 수 있다[3]. 그림 1은 자바로 구현된 컴포넌트의 Case 예이다. 'URL_Reader' 컴포넌트는 BufferedReader를 사용해서 URL을 직접 읽어오는 방법에 대한 소스이다. 만약, 어떤 사용자(프로그래머)가 BufferedReader를 사용해서 URL을 읽어오는 프로그램을 작성고자 할 때 프로그램 작성 방법을 모르거나 작성 패턴을 알고 싶을 때, 사용자는 'BufferedReader' 와 'URL'과 같이 라이브러리에 있

는 클래스 이름을 이용한 소스 코드를 그대로 질의함으로써 'URL_Reader' 컴포넌트를 검색할 수 있다. Case 기반 검색은 크게 2 단계로 이루어진다. 1 단계는 의미망(semantic net)을 구성하는 단계이다. 라이브러리에 저장된 클래스를 기반으로 한 컴포넌트는 파싱 과정을 거쳐 의미망을 구성한다. 의미망은 소스 코드의 'class' 등으로 구성된 노드와 'relevance' 등으로 구성된 간선으로 만들어 진다. 2 단계는 의미망을 이용한 검색 단계이다. 사용자가 질의로 준 소스 코드를 이용한 검색 단계이다. 질의 소스 코드에서 추출된 식별자가 의미망을 활성화시켜 연관된 컴포넌트를 검색한다.

Source Code

```
import java.net.*;
import java.io.*;

public class URL_Reader
{
   public static void main(String[] args)
   {
      URL kbs=new URL("http://www.kbs.co.kr");
      BufferedReader in=new BufferedReader(
         new InputStreamReader(kbs.openStream()));
      String inputLine;
      while ((inputLine=in.readLine())!=null)
         Sysem.out.println(inputLine);
      in.close();
   }
}
```

Goal Text

- *How to read directly from a URL using BufferedReader?*
- *How to copy directly from a URL to an output stream?*

그림 1. Case-URL_Reader

3. 의미망의 구축

3.1 의미망

Case 단위로 라이브러리에 저장된 각 컴포넌트는 파싱 단계를 거쳐 파싱 트리를 형성한다. 'class', 'interface', 'method', 그리고 'variable' 등을 식별자로 추출하였다. 추출 과정은 컴포넌트 소스 코드를 읽어 각 식별자를 추출하고 식별자 사이의 관계 정보를 저장한다. 의미망은 파싱 트리로부터 만들어진다. 파싱으로부터 생성된 'class', 'interface', 'method', 'variable'은 의미망에서 노드로 구성되며, 'relevance', 'subclass', 'implements', 'member', 그리고 'invoke'등은 클래스들 간의 관계로 취급하여 노드 간 간선으로 구성된다. 그림 1의 Case-URL_Reader에 대한 의미망은 그림 2와 같이 구성된다.

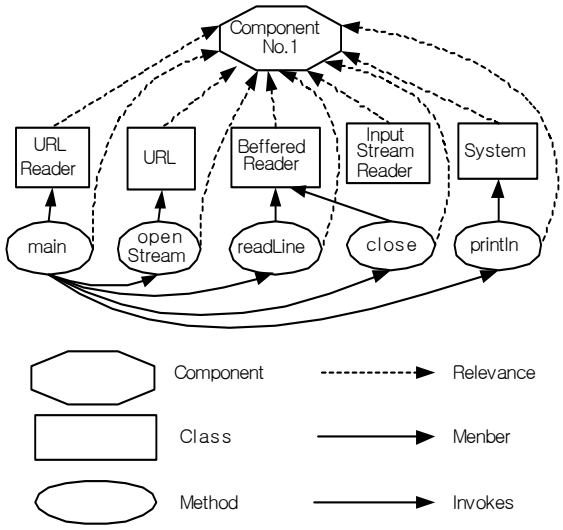

그림 2. 'URL_Reader' 컴포넌트의 의미망

컴포넌트 검색은 사용자에 의해 주어진 질의가 의미망을 활성화시킴으로써 이루어진다. 사용자는 자신이 작성하고 있는 소스 코드를 질의로 사용한다. 질의로 주워진 소스 코드는 구문 분석을 통해 식별자로 추출되는데, 사용자 소스 파일을 입력받아 먼저 토큰 단위로 식별자 이름을 모두 추출한 후, 처음 식별자명에서 다음 식별자명이 나타날 때까지 비교하면서 식별자 수만큼의 반복으로 정보를 추출한다. 추출된 식별자는 의미망에 있는 노드와 비교하기 위하여 inexact string matching algorithm 등을 이용한다[1].

소스 코드의 식별자는 각 컴포넌트의 의미망에 있는 노드와 비교되어 매치되는 노드를 활성화시킨다. 노드의 활성화는 spreading activation algorithm에 의해 이루어진다[4]. 의미망의 노드와 간선의 초기 활성값을 이용하여 서로 연결되어 있는 노드를 참조해 가면서 활성값을 계산하게 된다. 순환이 반복될수록 활성값은 안정되며 활성값이 기준에 미달되는 부분은 자동으로 제거되어 계산과정이 종료된다. 최종적으로 활성값이 가장 높은 컴포넌트들이 검색된다.

3.2 객체지향 시소러스 기반의 의미망 구축

본 연구에서 제안하는 의미망의 구축은 객체지향 시소러스를 기반으로 한다[5]. 이는 객체지향 코드에서 서로 관련 있는 클래스들을 일정한 기준에 따라 그룹화해서 여러 개의 클래스 그룹으로 구성하는 방식이다. 여기에서 여러 개의 클래스 그룹은 개념들 간에 나타나는 일종의 시소러스를 의미하며, 클래스

의 상속관계를 이용하여 개념들 사이의 관계를 자연스럽게 표현해야 한다.

다음은 본 연구에서 제안하는 객체지향 시소러스를 기반으로하여 의미망을 구축하는 단계이다.

▶ 1 단계 : 클래스의 개념적 분류
▶ 2 단계 : 클래스 범주 생성
▶ 3 단계 : 클래스와 범주의 관계값 계산
▶ 4 단계 : 객체지향 시소러스 생성
▶ 5 단계 : 시소러스 기반의 의미망 구축

그림 3은 소스 코드로부터 시소러스가 구축되는 전반적인 과정을 나타낸 것이다.

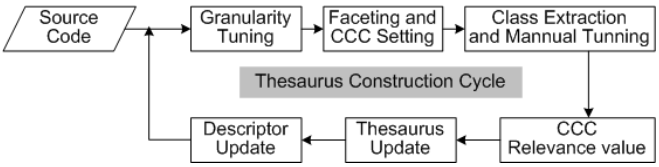

그림 3. 객체지향 시소러스 구축 과정

(1) 1 단계 : 클래스의 개념적 분류
클래스를 기능과 개념에 따라 여러 개로 나눈다. 이 과정은 도메인 전문가에 의해 행해지며 시스템의 응용에 따라 모든 클래스를 최적으로 표현할 수 있는 개념을 선정한다. 각 개념은 패싯 항목으로 표현되고 이에 따라 클래스가 분류되어진다.

(2) 2 단계 : 클래스 범주 생성
개념적으로 분류된 클래스를 이용하여 클래스 개념 범주를 생성한다. 클래스가 사용될 수 있는 여러 경험적 상황을 패싯 항목으로 설정하는 패싯 분류방법을 사용한다. 클래스의 사용범위가 한가지로 국한되지 않고 여러 경우가 가능하며 이에 따른 경험적 솔루션을 제시해준다는 점에서 컴포넌트의 분류와 검색을 위해 클래스를 이용한 패싯 분류는 효율적인 방법이다.

(3) 3 단계 : 클래스와 범주의 관계값 계산
분류된 범주를 사용하여 클래스와 범주간의 관계값을 계산한다. 이 과정은 범주를 이용한 클래스 빈도를 계산하는 과정이다. 이 빈도값이 범주별 클래스 관계값이 되며, 이 관계값에 의해 초기 시소러스 유의어 테이블이 구성된다.

(4) 4 단계 : 객체지향 시소러스 생성
클래스와 범주의 관계값을 이용하여 퍼지 시소러스를 구축한다. 각 클래스와 범주에 대한 매칭 정도를 비교함으로써 이들 사이의 퍼지 정도를 계산하여 시소러스를 구축한다.

(5) 5 단계 : 시소러스 기반의 의미망 구축
소스 코드로부터 추출된 각 클래스와 메소드는 의미망의 노드와 간선으로 표현되며, 초기 각 노드의 관련값은 시소러스의 유의값으로 설정된다.

4. 검색방법의 개선

4.1 의미망을 이용한 컴포넌트 검색

본 연구는 그림 4와 같이 객체 지향 소스 코드를 재사용하기 위해 의미망을 효율적인 구현하여 기능적으로 유사하거나 관련이 깊은 컴포넌트를 검색하는 것이다.

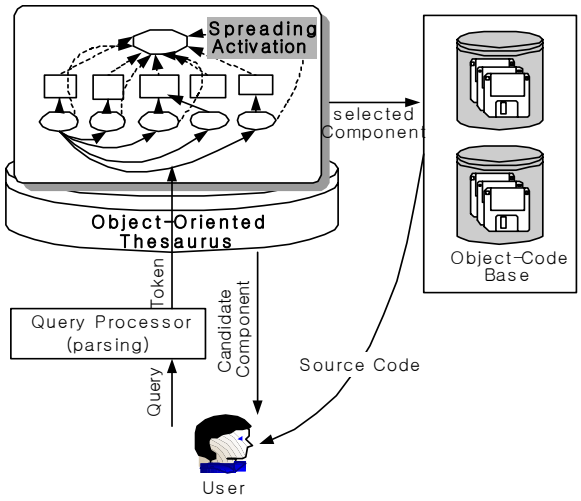

그림 4. 의미망을 이용한 컴포넌트 검색

사용자는 컴포넌트의 행위적 특성을 입력하거나, 작업 중인 소스 코드를 질의로 입력할 수 있다. 입력된 질의는 질의어 처리기를 통하여 구문 분석되어 각 Case 별 의미망과 비교된다. 의미망에 있는 클래스와 컴포넌트는 객체지향 시소러스에 의해 초기 관련값이 설정되어 있고, 파싱된 질의는 의미망에서 매칭되는 각 노드를 활성화시킨다. 의미망의 노드와 간선의 초기 활성값을 이용하여 서로 연결되어 있는 노드를 참조해 가면서 활성값을 계산하게 되고, 순환이 반복될수록 활성값은 안정되며 참조회수가 기준에 미달되는 부분은 자동으로 제거되어 계산과정이 종료된다. 최종적으로 활성값이 가장 높은 컴포넌트들이 검색된다. 후보 컴포넌트는 순위에 따라 사용자에게 제공된다. 사용자는 이들 중 한 컴포넌트를 선택하면 선택된 컴포넌트의 소스 코드가 제공된다.

4.2 검색 방법의 개선

spreading activation algorithm은 몇 가지 문제점을 지니고 있다. 이 알고리즘은 컴포넌트 간의 연

관성에 대한 연결강도(connection relaxation)를 기초로 한 방법이다. 이 기술은 활성값이 안정되거나, 사용자가 설정한 최대 사이클 수만큼 실행하여 검색어에 연관이 있지만 직접 연결되지 않은 유사한 컴포넌트도 검색하도록 해준다. 이처럼 spreading activation algorithm은 직접 인덱싱되어 있지 않은 컴포넌트까지 검색할 수 있는 효율적인 검색 방법이며 라이브러리에 컴포넌트들을 구축할 때 각 항목들을 일일이 인덱싱하지 않아도 되기 때문에 많은 비용이 절감된다. 그러나 검색할 때 활성값을 이용하여 유사도를 측정하기 때문에 시간이 많이 걸리는 단점이 있다. 이는 컴포넌트들이 증가할수록 활성값의 계산회수가 지수적으로 증가하기 때문에 컴포넌트들이 많을수록 검색에는 어려움이 발생한다. 따라서 의미망을 효율적으로 구축하기 위해서는 spreading activation의 성능은 최대한 유지하면서 검색 속도를 향상시킬 수 있는 방법이 요구된다.

이에 본 연구에서는 spreading activation 방법의 단점을 해결하기 위하여 인덱싱이 적은 컴포넌트의 연결정보를 제거함으로써 활성값 계산회수를 줄여 검색시간을 단축시키는 방법으로 개선하고자 한다. 즉, 순환과정이 일정 수준 반복된 후 기준에 미치지 못하는 질의어나 컴포넌트의 연결정보를 제거하여 연산에서 제외시킴으로써 질의어의 확장범위를 줄여 보다 관계가 깊은 컴포넌트만을 검색하도록 하는 것이다.

제안한 spreading activation 방법은 다음과 같다.

▶ 1 단계 : 의미망의 각 노드에 시소러스의 유의값을 초기 활성값으로 설정.

▶ 2 단계 : 질의와 매칭되는 의미망의 노드 활성화.

▶ 3 단계 : 활성화된 노드의 초기 활성값이 연결된 다른 노드로 이동.

▶ 4 단계 : 순환 반복.

▶ 5 단계 : 노드 간 연결의 참조 회수를 이용하여 제거 기준 설정.

▶ 6 단계 : 참조회수가 기준에 못 미치면 연결삭제.

▶ 7 단계 : 순환이 끝나면 질의와 직접 연결된 컴포넌트의 활성값과 유사한 활성값을 갖는 컴포넌트 검색.

5. 결론

Case 기반 컴포넌트 검색을 위해 효율적인 의미망을 구축하기 위해서는 객체지향 파라다임을 검색에 적합한 형태로 해석·적용하고, 기존의 시소러스에서 이용된 관계성을 클래스와 컴포넌트의 관계로 재정의함으로써 의미망의 구축과 검색을 위한 새로운 기본 구조가 만들어져야 한다. 또한 구축된 의미망에 효율적인 검색 메카니즘을 적용함으로서 원하는 컴포넌트를 쉽게 검색할 수 있을 것이다. 따라서 본 연구는 소스 코드를 재사용하기 위한 Case 기반 검색에 있어서 효율적인 검색 시스템을 구축하는 방법을 제안하였다. 본 연구에서의 의미망 구축은 크게 2가지 관점에서 재사용에 유용하다. 하나는 프로그래머의 관심을 라이브러리 내에 있는 컴포넌트로 유도하여 재사용성을 높일 수 있다. 이는 검색된 컴포넌트 내에 각 클래스를 하이퍼링크로 라이브러리 API 연결시키는 방법 등을 사용하여 재사용을 유도할 수 있다.

다른 하나는, 여러 다양한 클래스들이 포함된 전형적인 유형의 프로그래밍 패턴을 제공함으로써 프로그래머로 하여금 프로그램의 가이드라인으로 사용할 수 있도록 도움을 준다.

참고문헌

[1] Markus, G, Derek B, "Case-Based Reuse of Software Examplets," GWEM 2003 Proceedings of the Workshop on Experience Management, Apr. 2003.

[2] Aamodt, A. Plaza, P, "Case-Based Reasoning: Fundamental Issue, Methodological Variants, and System Approaches," Artificial Intelligence Communications, Vol. 7, No. 1, pp.39-59, 1994.

[3] Gomes, P., Pereira, F.C., Paiva, P., Seco, N., Carreiro, P., Ferriera, J.L. & Bento, C., "Using CBR for Automation of Software Design Patterns," Proceedings of the Sixth European Workshop on Case-Based Reasoning, LNAI 2416, Springer, pp.543-548, 2002.

[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, pp.425-472, Oct. 1999.