韓國 컴퓨터情報學會 論文誌
第12卷 第3號, 2007. 7.

2007-12-3-1-9

웹 어플리케이션 재구조화를 위한 클러스터링에 사용되는 결합도 메트릭

이 은 주*, 박 근 덕**

Coupling Metrics for Web Pages Clustering in Restructuring of Web Applications

Lee Eun Joo*, Park Geun Duk**

요 약

웹 어플리케이션의 복잡도는 증가하고 생명주기는 점차 짧아지는 추세이다. 따라서 웹 어플리케이션의 유연성과 확장성을 향상시키는 재구조화가 필요하며, 클러스터링을 통해 시스템을 이해하고 재구조화하는 접근법이 사용되고 있다. 본 논문에서는 웹 페이지들을 클러스터링하기 위한 결합도 메트릭을 제안한다. 이를 위하여 웹 어플리케이션 모델을 정의하였으며, 이 모델에는 웹 페이지 사이의 관계 유형 및 파라미터의 개수 정보가 포함된다. 이를 기반으로 결합도 메트릭은 웹 페이지 사이의 직접적인 연결 강도와 간접적인 연결 강도를 고려하여 정의되었다. 두 페이지 사이에 직접적인 관계가 다수 존재하고 파라미터의 개수가 많을수록 직접적인 연결 강도는 높아지며, 두 페이지가 각각 다른 페이지들에 대해 가지는 연결 패턴이 유사할수록 간접적 연결 강도는 높아진다. 제안한 메트릭을 메트릭 검증 프레임웍을 이용하여 검증하고, 예제에 적용하여 기존 메트릭과의 비교 분석을 통하여 기존 메트릭의 단점을 보완하였음을 보인다.

Abstract

Due to the increasing complexity and shorter life cycle of web applications, web applications need to be restructured to improve flexibility and extensibility. These days approaches are being used where systems are understood and restructured through clustering techniques. In this paper, the coupling metrics are proposed for clustering web pages more effectively. To achieve this, web application models are defined, where the relationship between web pages and the numbers of parameters are included. Considering direct and indirect coupling strength based on these models, coupling metrics are defined. The more direct relations between two pages and the more parameters they have, the stronger direct coupling is. The higher indirect connectivity strength between two pages is, the more similar the patterns of relationships among other web pages are. We verify the suggested metrics according to the well known verification framework and provide a case study to show that our metrics complements some existing metrics.

▸ Keyword : 웹 어플리케이션(Web Application), 결합도(Coupling), 재구조화(Restructuring), 클러스터링(Clustering),

• 제1저자 : 이은주 • 교신저자 : 박근덕
• 접수일 : 2007.4.16, 심사일 : 2007.4.18, 심사완료일 : 2007. 4.28.
*경북대학교 컴퓨터공학과 전임강사 **호서대학교 컴퓨터공학부 전임강사
※ 이 논문은 2006년도 경북대학교 학술진흥연구비에 의하여 연구되었습니다.

Ⅰ. 서 론

현재 웹 어플리케이션의 복잡도(complexity)는 점차 증가하는 추세이며, 웹 어플리케이션의 생명 주기도 짧아지고 있다. 이는 웹 어플리케이션 개발 시에 변하는 요구사항을 기존의 시스템에 보다 빠르게 적용할 수 있어야 하며, 결국 웹 어플리케이션에 대한 유지보수성이 중요한 요소가 됨을 뜻한다. 또한 빨리 출시하기 위하여, 체계적인 개발 방법론이나 분석 및 설계 기법에 따르기가 어려우므로 어플리케이션이 잘 구조화 되어 있지 않는 경우가 많다. 이 역시 이후의 유지보수에 좋지 않은 영향을 미치는 요소이다.

따라서 웹 어플리케이션을 유연하고 확장성 있도록 하기 위하여 재구조화(restructuring)하는 것이 필요하며, 이를 위하여 대상이 되는 웹 어플리케이션을 잘 이해하는 것이 중요한 문제가 되었다 [1].

규모가 큰 시스템을 이해하기 위하여 역공학적 접근법이 사용되었다 [2][3]. 역공학방식은 대부분 시각적으로 보여주는 그래픽 표현방식을 쓰는데, 규모가 커질수록 결과의 복잡도 역시 높아져서 효용성이 떨어진다. [1]

역공학에서의 그래픽 표현 방식을 보완하기 위하여 시스템을 응집력 있는 부분으로 나누는 웹 클러스터링(clustering) 방식이 제안되었다 [1][4]. 이들은 웹 어플리케이션을 노드와 링크를 가지는 모델로 변환하고, 가중치를 가지는 링크에 기반하여 정의한 메트릭을 이용하여 계층적으로 클러스터링을 하고 있다. 그러나 [4]의 경우 링크의 유사도만을 고려하므로, 실제 직접적으로 관련이 있는 노드 사이의 관계 긴밀도가 반영이 되지 않을 수 있다. 그리고 [1]에서는 노드 사이에 간접적인 관계가 반영되지 않고 있으며, 파라미터를 고려하지 않고 있다.

이 두 방법은 모두 생성되는 웹 페이지 클러스터들의 응집력이 높으며 결합력이 낮아야 함을 전제로 하고 있다. 응집도(cohesion)과 결합도(coupling)는 소프트웨어 설계에서의 주요 척도로서 과거 구조적 설계(structured design)에서부터 객체 지향 설계(object-oriented design), 컴포넌트 기반 설계(component-based design)에 이르기까지 두루 이용되는 소프트웨어 메트릭(metric)이다. 응집도는, 한 모듈 내의 구성 요소들이 얼마나 긴밀하게 연계되어 특정 업무를 수행하고 있는 지를 보여주는 척도이고, 결합도는 모듈들 사이의 연결 정도를 보여주는 척도이다. 일반적으로 응집도는 높을수록, 결합도는 낮을수록 좋은 설계라고 여겨진다. 다수의 연구에서, 어떤 모듈을 구성하는 내부 요소들의 결합도를 이용하여 응집도를 구하는 것을 가정하고 있고 [1][4][5], 웹 클러스터의 경우에도 이러한 접근법의 적용이 가능하다.

따라서 본 연구에서는 웹 페이지 클러스터링에 이용 가능한 웹 어플리케이션의 결합도 메트릭을 제안한다. 그리고 정의한 메트릭을 검증 프레임웍을 통해 이론적으로 검증하고 사례를 통하여 그 유용성을 보인다. 웹 어플리케이션의 기본 단위를 웹 페이지와 이들이 이루는 관계로 본다. 그리고 이전 연구들의 결합도 메트릭을 보완하여 페이지 사이의 직, 간접적 관계를 중심으로 결합도 메트릭을 정의하였다. 웹 페이지들을 보다 큰 단위로 클러스터링을 할 때, 정의한 결합도를 하나의 기준으로 활용할 수 있을 것이다.

본 논문의 구성은 다음과 같다. 2장에서는 웹 메트릭 및 웹 페이지 클러스터링에 대한 관련 연구를 기술한다. 3장에서는 본 연구에 이용된 시스템 모델과 메트릭을 정의하고, 4장에서 이론적, 실제적인 검증을 수행한다. 그리고 5장에서는 결론 및 향후 과제에 대하여 기술한다.

Ⅱ. 관련 연구

웹 페이지를 이해하고 재구조화하기 위하여, 웹 메트릭을 활용하여 클러스터링하는 연구가 수행되었다 [1][4].

Lucca등은 웹 페이지 사이의 결합도를 이용하여 클러스터링하는 방법을 제안한다 [1]. 페이지사이의 *link, redirect, submit* 관계를 대상으로 링크의 방향성을 고려하여 결합도를 구한다. 두 웹 클러스터 A,B 사이의 결합도를 다음 식 (2.1)과 같이 정의하고 있다 [1].

$$COP(A,B) = p_{A\to B} * p_{B\leftarrow A} + p_{B\to A} * p_{A\leftarrow B} \quad \cdots\cdots (2.1)$$

여기서

$$p_{X\to Y} = \sum_{type \in TYPE} N_{type}(X\to Y) \cdot w_{type}^{out}(X)$$

$TYPE = \{link, redirect, submit\}$

$N_{type}(X\to Y)$: X에서 Y로 가는 각 type에 해당되는 연결의 개수

$w_{type}^{out}(X)$: 페이지 X로부터 해당 type으로 다른 페이지로 나가는(outgoing) 연결의 개수를 전체 연결의 개수로 나눈 값이다.

$p_{X \leftarrow Y}$: $p_{X \rightarrow Y}$가 X에서 Y로 나가는(outgoing) 연결에 대한 값이라면, $p_{X \leftarrow Y}$는 Y로부터 X로 들어오는(incoming) 연결에 대한 값이다.

(2.1)은 두 웹 클러스터 A, B사이의 결합도를 구하는 식이다. 결합도는 A를 기준으로 B에서 A로 들어오는(incoming) 경우와 A에서 B로 나가는(outgoing) 경우, 그리고 B를 기준으로 마찬가지의 경우를 고려하며, 각각의 횟수와 해당 경우에서의 가중치를 곱한다. 이때, 링크의 상대적인 중요성을 고려하고 있는데, 이를테면 페이지 A가 어떤 페이지 B와 하나의 링크가 존재하는 경우를, A가 다른 B를 포함한 다른 페이지들과 여러 링크가 존재하는 경우보다 더 높은 가중치를 가지도록 하고 있다. 그러나 include나 프레임 로드(load) 관계는 반영하지 않고 있으며 파라미터에 대한 고려가 없다. 그리고 간접적인 관계를 반영하지 않고 있다. 즉, 페이지 A와 B가 둘 다 페이지 Z를 참조하고 있는 경우, A와 B도 간접적인 관계를 가진다고 볼 수 있는데 [6], 이런 경우에 대한 고려가 부족하다.

[4]에서는 웹 페이지 사이에 링크 기반 유사도(link-based similarity)를 이용한 응집도 및 결합도 메트릭을 이용하여 클러스터링을 수행한다 [4]. 응집도와 결합도는 모두 페이지 사이의 링크 유사도를 기반으로 정의되었으며, 유전 알고리즘을 적용하여 웹 어플리케이션 전체의 응집도를 높이고 결합도를 낮추는 방향으로 클러스터링이 진행된다. [1]와 유사하게, 웹 페이지사이의 관계를 link, submit, redirect, include, build, load 등, 모두 여섯 종류로 분류하고, 이들 사이의 관계 가중치를 활용하여 웹 페이지 사이의 유사도를 정의하였다. 두 페이지 u, v사이의 관계 가중치는 아래 식 (2.2)와 같이 정의된다.

$$W(u,v) = \begin{cases} nlink(u,v) * nparam(u,v) * w_{ln} \\ nparams(u,v) * \{w_{su} \text{ or } w_{re} \text{ or } w_{ld}\} \\ w_{bu} \text{ or } w_{in} \end{cases} \quad \cdots\cdots (2.2)$$

여기서

w_{ln}, w_{su}, w_{re}, w_{ld}, w_{bu}, w_{in}은 각각 link, submit, redirect, load, build, include에 대한 가중치이며 nparam(u, v)는 u와 v 사이의 파라미터 개수이다.

즉 두 페이지 사이의 가중치는 관계의 유형과 파라미터 개수, 단일 관계의 가중치 w_i로 정의되는데, link관계일 경우 link의 개수와 파라미터 개수, 그리고 link에 대한 가중치의 곱이며, submit, redirect, load의 경우는 파라미터 개수와 각 단일 관계 가중치의 곱이다. 그리고 build와 include의 경우에는 단일 관계 가중치 자체가 두 페이지 사이의 관계 가중치이다.

페이지 가중치는 build가 다른 관계들보다 페이지 사이의 관련이 가장 높으며, link와 load, include는 가장 낮다. [4]에서는 정의한 페이지 사이의 관계 가중치에는 두 가지 단점이 존재한다. 우선, 두 페이지 사이에 하나 이상의 관계가 있음을 고려하지 않고 있다. 또한 두 웹 페이지 사이에 파라미터가 없는 경우도 다수 존재하는데, 이런 경우 페이지 가중치가 0이 되므로 관계 반영이 제대로 되지 않을 수 있다.

결국 각 웹 페이지는 웹 어플리케이션 내의 모든 페이지들과 관계 가중치로 구성된 벡터를 가지게 된다. 결국 이러한 벡터 유사도를 기반으로, 웹 페이지 u와 v사이의 페이지 유사도 PS(u, v)가 정의된다. [4]에서는 웹 페이지들의 모임인 웹 클러스터에 대하여 응집도와 결합도를 다음 식 (2.3) 과 같이 정의하고 있다.

$$COH(C) = \frac{\sum_{u,v \in C} PS(u,v)}{(|C|*|C|)} \quad \cdots\cdots (2.3)$$

$$COP(C_i, C_j) = \sum_{u \in C_i} \sum_{v \in C_j} PS(u,v)$$

여기서 C, Ci, Cj는 웹 페이지들의 집합인 웹 클러스터이며 |C|는 클러스터 C 내에 존재하는 웹 페이지들의 개수이다. 웹 클러스터의 응집도는, 두 웹 클러스터 내에 존재하는 페이지들 사이의 페이지 유사도의 평균치이며 웹 클러스터 사이의 결합도는 두 웹 클러스터 내에 존재하는 페이지들 사이의 페이지 유사도의 총합이다.

그러나 직접적으로 관련이 있는 페이지 사이의 관계 긴밀도가 반영이 되지 않는다. 즉, 페이지 A가 페이지 B를 직접적으로 참조하고 있는 경우라 하더라도 A, B의 연결 패턴이 다르다면 유사도는 낮게 되고 결과적으로 결합도가 낮아지게 된다.

Liu등은 웹 어플리케이션 테스팅을 위한 모델을 제안하고 있다 [7]. 이 모델은 웹 페이지 사이의 request, redirect, response, navigation, inheritance, aggregation, association 관계를 정의한다. 여기서는 웹 페이지를 일종의 객체와 같이 취급하고 있으며 inheritance, aggregation, association관계는 일반 객체지향 프로그램에서의 관계와 같다. 요청(request), 응답(response)관계는 클라이언트-서버 환경에서의 테스팅을 고려한 관점이 반영된 것이다.

웹 어플리케이션이 대하여 결합도 이외에, 크기(size,), 재사용성(reusability), 복잡도(complexity), 노력(effort) 등에 관련하여 여러 메트릭들이 정의되었다 [8][9]. 여기서 웹 페이지 사이의 연결 관련하여 정의된 부분이 연결도(connectivity)인데, 이는 복잡도 영역에 속하는 것으로 전체 내부 링크의 개수를 이용한다. 또한 연결도의 평균, 즉 연결도를 전체 페이지 수로 나눈 것을 연결 밀도(connectivity density)로 정의하고 있다. 본 논문에서 결합도를 측정하기 위해 정의한 연결 강도(connectivity strength)는 페이지 사이의 상호 관계(inter-relationship)으로서, [8]의 연결도와는 차이가 있다.

III. 결합도 메트릭

3.1 모델 명세

본 절에서는 웹 어플리케이션 및 웹 클러스터 모델에 대한 명세를 기술한다. 이 명세는 3.2절에서 메트릭 정의에서 이용된다.

웹 어플리케이션은 웹 페이지를 나타내는 노드와 웹 페이지들 사이의 관계를 나타내는 에지로 구성된다.

[정의 1] 웹 어플리케이션 그래프 WAG

- WAG = 〈N, E〉

- N은 웹 페이지를 나타내는 노드(node)
- E는 웹 페이지 사이의 에지(edge)

WAG은 방향성이 있는 그래프(directed graph)로서, 노드의 집합 N과 에지의 집합 E로 구성된다.

[정의 2] 유한한 페이지의 집합 N

N은 유한한 웹 페이지들의 집합이다. 웹 페이지는 본 연구에서는 [4]에서와 같이 정적 페이지와 동적 페이지가 여기에 해당한다.

[정의 3] 페이지 사이의 유한한 에지의 집합 E

$E \subseteq N \times N$ 는 페이지 사이의 유한한 에지의 집합으로, 속성 $EDGE_E$를 갖는다. 이 속성을 이용하여 페이지 사이의 연결 강도를 구한다. 어떤 에지 E가 가지는 속성의 집합 $EDGE_E$는 다음과 같다.

- EDGEE={⟨etype, nparam, freq⟩| etype ∈ ETYPE, nparam은 0이상의 정수, freq는 1이상의 정수}

- etype은 해당 에지의 유형(ETYPE)을 나타내며, 아래 〈정의 4〉의 한 유형에 해당된다.
- nparam은 E의 해당 유형에서 발생하는 파라미터의 개수로서, 0개 이상의 정수 값을 가진다.
- freq는 어떤 두 노드 사이에 발생하는 타입과 파라미터가 같은 경우의 횟수이다. 어떤 두 노드 사이에 발생하는 타입은 하나 이상이 될 수 있고, 두 노드 사이에 연결 타입과 파라미터 개수가 동일한 것이 하나 이상 있을 수 있다. 예를 들어 노드 (A, B)를 연결하는 에지 E_i가 있다고 가정하자. 웹 페이지 A에서 B로 submit관계가 존재하고 파라미터 개수가 두 개이며, 또한 A에서 B로 파라미터 없는 하이퍼링크가 두 개 존재한다면 Edge(A, B)={⟨*submit*, 2, 1⟩, ⟨*link*, 0, 2⟩}이 될 것이다.

- edge(E)={⟨etype, nparam, freq⟩| etype ∈ ETYPE, nparam, freq ∈ integer}

edge(*E*)는 에지 E에 대해 기 정의한 속성의 집합을 추출해주는 함수이다.

[정의 4] E의 유형 ETYPE

- ETYPE={link, submit, include, load, redirect}

에지 $E \subseteq N \times N$ 의 종류는 다음과 같다.

$E = E_{ln} \cup E_{su} \cup E_{in} \cup E_{ld} \cup E_{re}$

- $E_{ln} \subseteq N \times N$는 방향성을 가지는 에지들의 집합으로서 한 노드에서 다른 노드로 하이퍼링크(hyperlink)가 있는 경우이다. 예를 들어 $(v1, v2) \in E_{ln}$은 v1에서 v2로 가는 하이퍼링크가 존재함을 뜻하며, 이러한 관계를 *link*라고 한다.
- $E_{su} \subseteq N \times N$는 방향성을 가지는 에지들의 집합으로서 한 노드에서 다른 노드로 폼(form)을 통하여 데이터를 보내는 관계이다. 예를 들어 $(v1, v2) \in E_{su}$은 v1에서 폼을 이용하여 v2로 데이터를 보내는 경우이다. 이러한 관계를 *submit*관계라고 한다.
- $E_{in} \subseteq N \times N$는 방향성을 가지는 에지들의 집합으로서 한 노드가 다른 노드를 포함하는(include) 관계이다. 예를 들어 $(v1, v2) \in E_{in}$은 v1이 v2를 포함하는 것을 말한다. 이러한 관계를 *include*관계라고 한다.
- $E_{ld} \subseteq N \times N$는 방향성을 가지는 에지들의 집합으로서 한 노드가 다른 노드를 로드하는(load) 관계로 프레임(frame) 구조가 이에 해당한다. 예를 들어 $(v1, v2) \in E_{in}$은 v1노드 내부 한 프레임으로 v2를 포함하는 것을 말한다. 이러한 관계를 *load*관계라고 한다.
- $E_{re} \subseteq N \times N$는 방향성을 가지는 에지들의 집합으로서 한 노드에서 다른 노드로 이동(redirect)하는 관계이다. 예를 들어 $(v1, v2) \in E_{re}$은 v1에서 v2로 자동 이

동하는 것이다. 이러한 관계를 *redirect*관계라고 한다.

• edgeType(E)={etype| etype ∈ ETYPE}

edgeType(E)는 에지 E가 가지는 관계 유형을 추출해 주는 함수이다.

[1][4]에서는 2장에서 소개한 바와 같이 웹 페이지 사이의 여섯 종류의 관계를 소개하고 있으나, 관계 중에서 build의 경우는 어떤 페이지가 동적으로 생성된 결과 페이지로서 실제 존재하는 페이지가 아니므로 클러스터 대상이 될 수 없다. 따라서 본 연구에서는 build의 경우는 제외하기로 한다.

<그림 1>은 웹 페이지 사이의 관계를 UML 표기법으로 나타낸 것이다 [1][4].

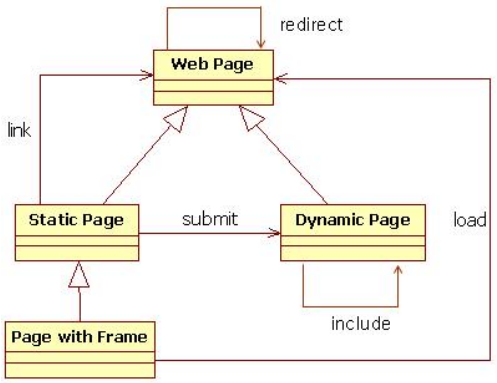

그림 1. 웹 페이지 관계
Fig 1. Relationships between Web Pages

[정의 5] 웹 클러스터 그래프 WCG

• WCG=〈N′, E′〉

- N′는 웹 클러스터를 나타내는 노드
- E′는 웹 클러스터 사이의 에지

[정의 6] 유한한 웹 클러스터의 집합 N′

N′는 유한한 웹 클러스터의 집합이다. 웹 클러스터는 웹 페이지들의 집합이며, 어떤 웹 클러스터 C는 다음과 같이 정의된다.

• C = { p | p ∈ N}

여기서, 페이지 사이의 관계는 이전 웹 어플리케이션에서와 마찬가지로 유지된다.

[정의 7] 웹 클러스터 사이의 유한한 에지의 집합 E′

E′⊆N′×N′는 가중치를 가지는 유한한 에지의 집합으로 속성 $EDGE_{E'}$를 갖는다.

• EDGEE′={〈ctype〉| ctype ∈ CTYPE}

[정의 8] E′의 유형 CTYPE

본 연구에서는 웹 클러스터 사이의 관계 유형인 CTYPE에서 방향성을 가지는 연관 관계(association relationship)를 정의한다.

- E′ ⊆N′×N′는 방향성을 가지는 에지의 집합으로, 어떤 웹 클러스터 C_1과 C_2에 대하여, E′=(C_1, C_2)라 하자. C_1내의 한 페이지에서 C_2내의 어떤 페이지로 ETYPE에 해당하는 관계가 존재할 경우, C_1에서 C_2로 연관 관계를 가진다고 한다.

• outAssoc(C)={Cj | (∃pi∈C) ∧Cj∈(N′-C)∧(∃pj∈Cj)∧edgeType(pi, pj)≠∅}

outAssoc(C)는 클러스터 C에서 연관 관계를 가지고 있는 클러스터들의 집합이다.

• *inAssoc*(*C*)={C_j | (∃p_i∈C) ∧C_j∈(N′-C)∧(∃p_j∈C_j)∧edgeType(p_j, p_i)≠∅}

inAssoc(C)는 클러스터 C로 연관 관계를 가지고 있는 클러스터들의 집합이다.

3.2 결합도

본 절에서는 웹 클러스터의 결합도를 측정하기 위한 메트릭에 대하여 정의한다. 메트릭 정의 시, 페이지 사이의 직접적인 관계와 간접적인 관계를 모두 고려하도록 한다.

페이지 사이의 직접적인 관계란, ETYPE에 속하는 관계가 존재함을 말한다. 실제, 도큐먼트(document) 사이에 링크가 존재한다는 것은 두 도큐먼트 사이에 의미적인 관계(semantic relation)이 존재하며, 이들 사이의 최단 경로가 멀면 멀수록 의미적으로 관련이 약해짐을 뜻한다 [6]. 따라서 어떤 페이지 A와 B사이에 직접적인 어떤 관계가 존재한다면 결합도는 증가해야 할 것이다. 그리고 어떤 두 모듈 사이의 연결 강도를 측정할 때, 연결 시 교환되는 파라미터 역시 고려해야 한다 [4][10]. [10]에서는 파라미터의 개수와 자료형을 고려하고 있으나, 웹 어플리케이션에서는 데이터의 자료형에 대한 구분이 불명확하므로 본 논문에서는 파라미터의 개수만 고려하기로 한다.

페이지 A와 B가 동일한 페이지 Z를 참조한다면 이것은 A와 B가 간접적으로 관계가 있다고 볼 수 있다. 이는 인용분석(citation analysis)으로부터 도출된 연결 기반의 유사도 메트릭을 응용한 것이다 [6]. 이는 과거의 참조 결합(bibliographically coupled) 개념으로부터 나온 것으로 두 도큐먼트 i와 j가 공통으로 참조하는 도큐먼트의 개수에

비례하도록 정의하였다 [11].

본 논문에서는 이러한 관계를 모두 반영하기 위하여, 개별 페이지 사이의 직·간접적 연결 강도(connectivity strength) 메트릭을 정의하고 이 연결 강도 메트릭을 이용하여 웹 페이지 사이의 연결 강도를 정의한다. 그리고 웹 클러스터의 결합도는 이 웹 페이지 사이의 연결 강도를 이용하여 정의한다.

[정의 9] 웹 페이지 사이의 직접적 연결 강도

웹 페이지 사이의 다섯 가지 관계가 정의되었고, 각 관계에 대하여 경우에 따라 전달되는 파라미터가 존재할 수 있다. 웹 페이지 i, j 사이의 직접적 연결 강도 DCS(i, j)는 아래 식 (3.1)과 같다.

$$DCS(i,j) = \sum_{<t,n,f> \in edge(E)} w_t \bullet (n+1) \bullet f \quad \cdots\cdots (3.1)$$

여기서

- E={(i, j), (j, i)}

에지는 방향성을 가지므로 두 웹 페이지 사이의 연결 강도를 구하기 위하여 (i, j)와 (j, i)의 경우를 모두 고려한다.

- w_t는 연결유형 t가 가지는 가중치이며 0부터 1사이의 값을 갖는다. Lucca등은 경험치로서 submit의 가중치가 가장 높고, redirect가 그 다음, link의 가중치가 가장 낮다고 주장한다 [1]. [4]에서도 유사한 방식으로 가중치를 정의하고 있다. t를 연결유형으로 보고, w_t를 t에 대한 가중치라고 볼 때, 본 논문에서는 웹 페이지 사이의 가중치를 다음과 같이 설정하도록 한다.

$w_{su}+w_{li}+w_{in}+w_{ld}+w_{re}=1.0,\ w_{li},\ w_{in},\ w_{ld} \le w_{re} \le w_{su}$

[4]에서는 파라미터의 개수를 고려하여 메트릭을 정의하였으나, 파라미터가 없는 경우 전체 식의 값이 0이 된다. 이는, 두 페이지 사이에 하이퍼링크가 여러 개 존재하더라도 전달하는 파라미터가 없는 경우 페이지 사이의 연결 가중치가 0이 되므로 연결 관계를 정확히 반영하기 어렵다. 따라서 본 식에서는 파라미터가 없는 경우를 1로 정하였다.

[정의 10] 웹 페이지 사이의 간접적 연결 강도

간접적 연결 강도에서의 가정은, 공통된 페이지를 참조하는 두 페이지의 경우 간접적으로 연관이 되어 있다는 것이다. 본 연구에서는 각 페이지의 팬아웃(fan-out) 연결을 대상으로 삼는다. 각 노드 i는 다음과 같은 연결 가중치 벡터를 가진다.

V(i)=〈w_1, w_2, …, w_n〉, n은 전체 노드의 개수.

여기서

$$w_j = \sum_{<t,n,f> \in edge(i,j)} w(t) \bullet f$$

i와 j의 연결 가중치는, i와 j사이의 관계 가중치와 해당 관계의 개수를 곱하여 구한다. 여기서 파라미터의 개수를 배제한 이유는, 연결의 유사도를 고려하는 것이 목적인데, 파라미터의 개수를 적용하는 것이 결과를 왜곡할 가능성이 높기 때문이다.

두 페이지 i, j의 간접적 연결 강도 ICS(i, j)는 벡터 공간에서의 두 벡터의 유사도를 측정하는 식을 이용하며 다음 식 (3.2)와 같이 계산한다.

$$ICS(i,j) = \begin{cases} 0 & if\ V(i) \text{ or } V(j) = 0 \\ \dfrac{V(i) \bullet V(j)}{|V(i)||V(j)|} & otherwise \end{cases} \quad \cdots\cdots (3.2)$$

ICS(i, j)는 두 가중치 벡터가 이루는 각의 코사인 값이므로 0부터 1사이의 값을 가진다. ICS(i, j)의 최대값은, V(i)와 V(j)가 같은 경우이다. 즉, 두 페이지 i와 j가 다른 페이지와 가지는 연결 가중치가 모두 같은 경우 최대값 1을 갖는다. 또한 두 벡터 중 하나가 영벡터이거나, 즉 다른 페이지와 어떠한 관계도 가지고 있지 않거나, 두 벡터가 관계를 가지고 있는 페이지들의 집합이 공집합이 되는 경우, ICS(i, j)는 최소값 0을 갖는다.

[정의 11] 웹 페이지 사이의 연결 강도

ICS(i, j)와 DCS(i, j)를 이용하여 두 웹 페이지 i, j 사이의 연결 강도 WCS(i, j)를 다음 식 (3.3)과 같이 정의한다.

$$WCS(i,j) = (1 + ICS(i,j)) \bullet DCS(i,j) \quad \cdots\cdots (3.3)$$

본 논문에서는 두 페이지가 직접적으로 연관이 있는 경우가 간접적으로 연관이 있는 경우보다 더 높은 결합도를 가진다고 가정한다. 이를테면 A가 B에게 "이름"과 "주민번호"를 파라미터로 전달하는 *submit*관계와, X와 Y가 같은 페이지 Z를 참조하는 관계가 있다고 하자. 대부분의 경우 A가 전달한 파라미터들을 B가 처리하고 결과를 반영하게 되며, 파라미터를 매개 하여 A와 B가 강하게 결합되어 있음을 뜻한다. 그러나 X와 Y가 특정 페이지를 참조하는 경우를 생각해 보면, 과거 도큐먼트의 레퍼런스와 같은 경우는 X, Y, Z가 어느 정도 의미적 관련이 있을 가능성이 높겠지만 여러 웹 페이지 상의 관계에 있어

서는 그들 사이에 항상 어떤 의미적 관계가 있다고 보기 힘들다. 이를테면, 홈(home)으로 가는 링크가 존재하는 많은 페이지들이나, 헤더(header)나 푸터(footer)를 포함(*include*)하는 페이지들이 모두 서로 어떤 의미적 관계를 가진다고 보기 어렵다. 실제로 어떤 한 단위로 클러스터링되어야 하는 경우는 전자의 경우일 것이다. 따라서 본 연구에서는 간접적 연결 강도를 직접적 연결 강도의 일종의 가중치 역할을 하도록 정의하였다.

〔정의 12〕 웹 클러스터 사이의 연결 강도

두 웹 클러스터 C_i와 C_j사이의 연결 강도는 C_i에 속하는 페이지와 C_j에 속하는 페이지 사이의 연결 강도의 합으로 정의되며 다음 식 (3.4)와 같다.

$$CCS(C_i, C_j) = \sum_{p \in C_i} \sum_{q \in C_j} WCS(p, q) \quad \cdots\cdots (3.4)$$

〔정의 13〕 웹 클러스터의 결합도

어떤 웹 클러스터 C의 결합도는, C와 다른 클러스터들 사이의 연결 강도의 총합으로 구하며 아래 식 (3.5)와 같다.

$$CCOP(C) = \sum_{C_i \in ASSOC_C} CCS(C, C_i) \quad \cdots\cdots (3.5)$$

여기서

$ASSOC_C = inAssoc(C) \cup outAssoc(C)$

Ⅳ. 검증

4.1 이론적 검증

Briand의 검증 프레임웍〔12〕은 수학적 개념에 기반한 메트릭 검증 프레임웍으로서 크기(size), 길이(length), 복잡도(complexity), 응집도, 결합도 메트릭이 가져야 하는 속성을 정의한다. 이 프레임웍으로 새로운 메트릭을 정의하기 위한 기준을 삼을 수 있다. 본 연구에서 Briand 프레임웍에서의 모듈에 해당되는 단위를 웹 클러스터로 간주하며, 웹 클러스터 C의 결합도 CCOP(C)에 대하여 각 속성이 만족함을 보인다.

- 결합도 속성1: 비음수성(Nonnegativity)

CCOP(C)는 음수가 아니다

(증명) C와 다른 모든 클러스터 사이에 관계가 없을 경우 CCOP(C)는 0이 된다. 만일 C와 어떤 클러스터 C_i 사이에 연관 관계가 존재한다고 하자. 이것은 C 내의 한 웹 페이지 p와 C_i 내의 웹 페이지 p_i사이에 어떤 관계 R∈ETYPE이 존재함을 의미한다. 따라서 DCS(p, p_i)가 계산되어, 최종적으로 WCS(p, p_i)가 CCOP(C)에 더해진다. 따라서 어떠한 경우에도 CCOP(C)는 음수를 가질 수 없다.

- 결합도 속성2: 널 값(Null value)

C와 다른 클러스터 사이에 관계가 없으면 CCOP(C)는 널 값을 갖는다

(증명) C와 다른 클러스터 사이에 관계가 없으면, 내부적으로 C와 다른 클러스터들 사이에 어떠한 관계 E∈ETYPE도 존재하지 않음을 뜻한다. 이는 결국 C내부의 어떠한 웹 페이지 p에 대하여 DCS(p, p_i) (p_i는 C를 제외한, 다른 클러스터 내의 페이지)가 모두 0임을 뜻한다. 따라서 WCS(p, p_i)는 0이 되어 C의 결합도는 0이 된다.

- 결합도 속성3: 단조성(Monotonicity)

C와 다른 클러스터들 간에 관계가 추가되면 CCOP(C)는 감소하지 않는다.

(증명) C와 다른 클러스터 C′사이에 연관 관계가 추가되었다는 것은, C 내의 한 페이지 p와 C′내의 한 페이지 p′사이에 어떤 관계 R∈ETYPE이 추가되었음을 뜻한다. 이 경우 DCS(p, p′)가 C의 결합도 CCOP(C)에 더해진다. 그러나 본 연구에서 간접적인 관계도 고려하고 있으므로, p의 가중치 벡터와 p′의 가중치 벡터 사이의 유사도의 증감을 고려해야 한다. 이 가중치 벡터 사이의 유사도는 일부 경우에 따라 감소할 수 있으므로 DCS(p, p′)가 증가하더라도 전체 결합도 CCOP(C)가 감소할 가능성이 존재한다.

- 결합도 속성4: 클러스터 병합(Merging of Clusters)

두 클러스터가 병합된 후의 클러스터의 결합도는, 각 클러스터 결합도의 합보다 크지 않다.

(증명) 두 클러스터 C_1, C_2가 병합한 클러스터를 $C_{1\cup2}$라고 하자. $C_{1\cup2}$의 결합도 CCOP($C_{1\cup2}$)는 C_1의 결합도 CCOP(C_1)와 C_2의 결합도 CCOP(C_2)에서 CCS(C_1, C_2)를 뺀 값과 같다. 따라서 병합된 클러스터의 결합도는 각 클러스터의 결합도 합보다 크지 않다.

- 결합도 속성5: 서로 소인 클러스터 병합(Disjoint Cluster Additivity)

서로 소인 두 클러스터를 병합한 클러스터의 결합도는 각 클러스터 결합도의 합과 같다.

(증명) 서로 소인 두 클러스터는, 클러스터 사이에 연관

관계가 없는 클러스터를 말한다. 서로 소인 두 클러스터 C_1, C_2가 있다고 가정하자. 두 클러스터 사이에 연관 관계가 없으므로 $CCS(C_1, C_2)$는 0이 된다. 따라서 〈결합도 속성 4〉에서 기술한 바와 같이 $CCOP(C_{1\cup2})$는 $CCOP(C_1)$와 $CCOP(C_2)$의 합에서 $CCS(C_1, C_2)$를 뺀 값과 같은데 $CCS(C_1, C_2)$가 0이므로, $CCOP(C_{1\cup2})$는 두 클러스터의 결합도를 더한 값과 같다.

본 논문에서 정의한 결합도는 Briand가 제안한 결합도 속성 중에서 〈결합도 속성3〉을 제외한 다른 속성들을 만족한다. 〈결합도 속성3〉은, 직접적 연결 강도만을 고려한다면 만족한다. 그러나 간접적 연결 강도를 부수적인 가중치로 채택하였으므로, 관계가 추가되는 것이 일부 경우에 간접적 연결 강도를 감소시키고, 감소 폭에 따라 전체 결합도가 감소하는 경우가 존재한다. 즉, 이것은 결합도의 의미를 직접적인 연결 관계 이외에 간접적인 연결을 고려하여 발생한 결과이다.

4.2 사례 연구

본 절에서는 [1]에서 이용하고 있는 예제를 통하여 기 정의된 웹 결합도 [4], [1]와 본 논문에서 정의한 메트릭을 적용한 결과를 보이고 두 방법을 비교분석한다.

웹 클러스터의 결합도는, 웹 클러스터 내부 페이지들과, 외부 클러스터 내의 페이지들의 연결 강도의 합으로 정의 가능하므로 본 절에서의 대상은 웹 페이지 사이의 연결 강도로 한정짓는다. 본 연구에서의 웹 페이지 연결 강도는, [4]의 페이지 유사도와 [1]의 페이지 결합도에 비견될 수 있다. 왜냐하면, 이들을 이용하여 웹 클러스터의 결합도를 계산하고 있기 때문이다.

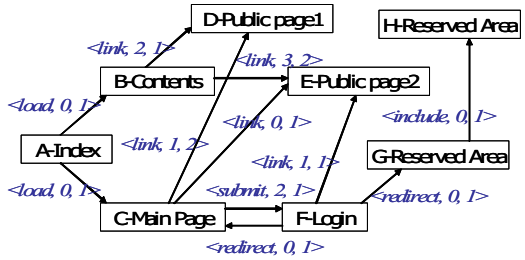

그림 2. 웹 어플리케이션 예제
Fig 2. Web Application Example

예제(〈그림 2〉)는 [1]의 예제에 파라미터와 include관계를 추가, 활용하였으며, 전기한 바와 같이 build관계는 제외하였다.

웹 페이지 사이의 관계 가중치는 〈정의 9〉에 따라 다음 〈표 1〉과 같이 정하였다.

표 1. 웹 페이지 사이의 관계 가중치
Table 1. Relationship Weights between Web Pages

wsu	wli	win	wld	wre
0.3	0.15	0.15	0.15	0.25

아래 〈그림 3〉은 본 논문에서 정의한 메트릭을 적용한 웹 페이지 사이의 연결 강도를 나타낸다.

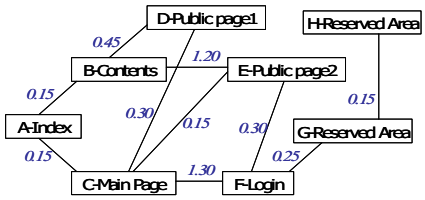

그림 3. 웹 페이지 연결 강도
Fig 3. Connectivity Strength between Web Pages

[4]에서 정의한 메트릭을 적용한 결과, 본 논문에서의 웹 페이지 연결 강도에 해당되는 페이지 사이의 유사도가 모두 0이 나왔다. 이것은, 2장 관련연구에서 기술한대로 파라미터가 없는 경우 *include* 관계를 제외하면 가중치가 0이 되므로 발생한 결과이다. 이를 조정하기 위하여 본 논문에서와 같은 방식으로 파라미터가 없는 경우를 1로 하고 적용해보면, C와 F사이에 0.025의 연결 유사도가 존재한다. 파라미터 적용을 조정할 때, 서로 직접적으로 연관이 있는 경우에도, 이들의 참조 패턴이 다르다면 페이지 사이의 유사도가 0이 되어 이후 클러스터링에서의 대상에서 제외되게 된다. 즉, 아무리 두 페이지가 여러 파라미터를 주고받으며 많은 링크로 연결되어 있다 하더라도 두 페이지가 다양한 관계로 연결되어 있는 페이지들의 집합이 다르다면, 유사도는 0이 된다. 따라서 [4]의 접근법은 다양한 관계와 파라미터로 연관되어 있는 현재 웹 어플리케이션에 적용하기에는 부족하다.

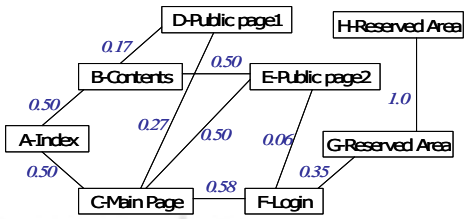

그림 4. 웹 페이지 결합도 [1]
Fig 4. Web Pages Coupling [1]

〈그림 4〉는 〈그림 2〉의 예제를 [1]에 적용한 결과이다. 여기서, Lucca02에서 제외된 include와 load관계에 대하여 본 논문에서 정의한 대로 link와 같은 가중치를 부여하였다.

결합도 메트릭의 경우 응집도와는 달리 정규화가 되어 있지 않으므로, 다르게 정의된 두 결합도의 절대적인 값을 비교하는 것은 큰 의미가 없다. 여기서는 해당 웹 페이지들 내에서 연결 강도(혹은 결합도)의 크기를 비교하는 것이 의미가 있다.

Lucca의 방법에서는, 페이지 내의 링크의 상대적인 개수까지 반영을 하므로, 한 페이지하고만 관련이 있을 경우, 즉 〈그림 4〉의 H와 G의 경우, 결합도가 높게 나타난다. 〈그림 3〉과 〈그림 4〉에서 페이지 C와 F의 결합도는 전체에서 각각 첫번째, 두번째로 높다. 본 예제에서 C와 F는 직, 간접적으로 연관되어 있다. 즉 각각 submit, redirect 관계로 직접적인 연결 강도를 가지며, 모두 E에 link관계를 가지므로 간접적인 연결 강도가 존재한다. 본 연구에서는 간접적 연결 강도가 가중치로 사용되었으므로, C와 F의 결합도가 가장 높게 나타났다. B와 E의 경우 예제 내에서의 결합도의 위치가 비슷하다. 그런데 〈그림 4〉에서 E와 F의 결합도가 가장 낮은데 반해 〈그림 3〉에서는 10개의 연결 중 네번째로 높다. 그 이유는, 본 연구(〈그림 3〉)에서는, E와 F가 파라미터가 존재하는 link 관계인데, 본 연구에서는 파라미터의 개수까지 반영을 하였기 때문이다. 그런데 [1]에서는 E나 F의 in/out 페이지 수가 많을수록 상대적으로 E와 F에 대한 연결의 가중치가 낮아져 전체 결합도가 낮아지기 때문이다.

[1]에서 소개된 예제에 대하여, 본 논문에서 정의한 메트릭과, [1], [4]에서 정의한 메트릭을 적용하고 결과를 비교해 보았다. [4]의 경우에는 파라미터 관련 문제를 보정하더라도, 직접적인 관계를 반영하기가 어렵다는 단점이 존재하였다. [1]의 경우, 어떤 페이지가 "index" 파일처럼 많은 링크를 가지고 있을 경우, 해당 페이지의 결합도는 낮아질 확률이 크다. 다시 말해, 어떤 페이지에 연관된 페이지가 많을 경우, 이 페이지와 연관된 페이지들의 결합도는 낮아지고, 〈그림 4〉의 H처럼 링크 개수가 작을 경우에는 결합도가 높아진다. 링크 개수가 작은 경우는 향후 클러스터링의 우선 순위가 되는 것이 의미가 있으나, 반대의 경우에는 의미를 두기가 어렵다. 또한 결합도에서 고려되는 파라미터 개수를 고려하고 있지 않다.

〈그림 3〉, 〈그림 4〉에서의 G와 H의 경우, H는 G하고만 관련이 있으므로 이후 H는 G와 같은 클러스터에 속하는 것이 타당하다. 이를 위하여 본 연구에서는 클러스터링 시 [10]에서와 같이, 클러스터링 시에 한 페이지하고만 관련이 있는 페이지는 하나의 클러스터에 두는 것을 일종의 휴리스틱으로 이용하여 보완해야 할 것이다.

V. 결론 및 향후 과제

본 논문에서는 웹 어플리케이션을 재구조화하기 위하여 이용되는 클러스터링 기법에서 클러스터링의 기준으로 활용 가능한 웹 페이지에서의 결합도 메트릭에 대하여 정의하였다. 여기에는, 웹 페이지 사이의 직접적인 연결 강도와 간접적인 연결 강도를 반영한 페이지 사이의 연결 강도와, 웹 클러스터 사이의 연결 강도 및 웹 클러스터의 결합도가 포함된다. 정의한 웹 클러스터의 결합도에 대하여 이론적 검증을 수행하여, 대부분의 속성을 만족함을 보였다. 또한 웹 클러스터 결합도를 구하기 위하여 이용되는 웹 페이지 사이의 연결 강도 메트릭에 대하여, 사례 연구를 통하여 기존 연구[1][4]과 본 연구의 결과를 비교분석하였다.

본 논문에서 정의한 결합도는, 이전 연구들의 단점을 보완하여 넘겨지는 파라미터와, 웹 페이지 사이의 직간접적 관계를 모두 고려하였다는 데에 의의가 있으며, 향후 클러스터링 기법에서 하나의 기준으로 활용할 수 있을 것이다. 그러나, [3]의 상대적인 결합도에 대한 고려가 부족하므로, 4.2절에서 기술한 바와 같이 이후 클러스터링에서 하나의 휴리스틱으로 보완을 해야 할 것이다.

향후 과제로서 본 논문에서 정의한 결합도를 이용하여, 실제 클러스터링 기법에 적용하여 그 유용성을 입증할 것이다.

참고문헌

[1] G. A. Di Lucca, A. R. Fasolino, F. Pace, P. Tramontana, U. De Carlini, "Comprehending Web Applications by a Clustering Based Approach," in Proc. of the International Workshop on Program Comprehension, pp.261-270, 2002

[2] F. Ricca and P. Tonella, "Understanding and Restructuring Web Sites with ReWeb," IEEE Multimedia, 2001.

[3] G. A. Di Lucca, M. Di Penta, and G. Antoniol, G. Casazza, "An approach for reverse engineering of web-based application," in Proc. of Working Conference on Reverse Engineering, pp. 231-240, 2001.

〔4〕 B.J. Lee, E. J. Lee, and C. S. Wu, "Genetic Algorithm Based Restructuring of Web Applications Using Web Page Relationships and Metrics," Lecture Notes in Computer Science, Springer-Verlag, Vol. 4113, pp.697-702, 2006.
〔5〕 T. Vernazza, G. Granatella, G. Succi, L. Benedicenti, and M. Mintchev, "Defining Metrics for Software Components," In Proc. of International Conference on Information Systems, Analysis, and Synthesis, 2000.
〔6〕 D. Dhyani, W.K. Ng, and S. S. Bhowmick, "A Survey of Web Metrics," ACM Computing Surveys, Vol. 34, No. 4, pp. 469-503, 2002.
〔7〕 C. Liu, D. C. Kung, P. Hsia, and C. Hsu, "Structural Testing of Web Applications," in Proc. of 11th International Symposium on Software Reliability Engineering, pp. 84-96, 2000.
〔8〕 R. Fewster and E. Mendes, "Measurement, Prediction and Risk Analysis for Web Applications," in Proc. of Software Metrics Symposium, pp. 338-348, 2001.
〔9〕 E. Mendes, N. Mosley, and S. Counsell, "Comparison of Web Size Measures for Predicting Web design and Authoring Effort," IEE Proceedings-Software 149(3), pp. 86-92, 2002.
〔10〕 이은주, "객체지향 시스템으로부터 컴포넌트를 식별하기 위한 모델 기반의 정량적 재공학," 정보처리학회 논문지 D, Vol.14-D, pp.67-82, 2007.
〔11〕 L. Egghe and R. Rousseau, Introduction to Informetrics, Elsevier Science Publishers, 1990.
〔12〕 L. C. Briand, S. Morasca, and V. R. Basili, "Property-Based Software Engineering Measurement," IEEE Transaction on Software Engineering, Vol. 22, No. 1, pp. 68-86, 1996.

저 자 소 개

이 은 주

2005년 2월 : 서울대학교 전기컴퓨터공학부 (공학박사)
2006년 3월~ 현재 : 경북대학교 컴퓨터공학과 전임강사
관심분야 : 웹공학, 재공학, 테스팅, 메트릭

박 근 덕

2005년 8월 : 서울대학교 전기컴퓨터공학부 (공학박사)
2006년 3월~ 현재 : 호서대학교 컴퓨터공학과 전임강사
관심분야 : 웹공학, 서비스 지향 컴퓨팅, XML 응용