2005 한국경영과학회/대한산업공학회 춘계공동학술대회
2005년 5월 13일~14일, 충북대학교

웹 사이트 구조 최적화를 위한 네트워크 모델링 접근법

이우기[1], 신광섭[2], 강석호[2], 김훈태[3]

[1]성결대학교 공과대학 컴퓨터학부

wook@sungkyul.edu

[2]서울대학교 공과대학 산업공학과

{whatever@ara, shkang@cybernet}snu.ac.kr

대진대학교 산업공학과

hoontae@daejin.ac.kr

Network Modelling Approach for Web Site Structure Optimization

Wookey Lee[1]
Dept. of Computer Science Sungkyul University
Kwangsup Shin[2], Sukho Kang[2],
Dept. of Industrial Engineering, Seoul National University
Hoontai Kim[3]
Dept. of Industrial Engineering, DaeJin University

Abstract

정보 통신 기술의 발달로 엄청난 양의 정보가 World Wide Web을 통해 저장되고 공유된다. 웹 정보의 양이 커질수록 이의 구조화 노력은 점증될 수밖에 없다. 본 논문의 목적은 웹을 유향그래프로 인식하고, 특히 웹 사이트에 초점을 맞추어 웹의 시작페이지(default page)와 이를 제외한 모든 페이지에 대해 최적구조화를 수행하되, 각 개별 웹 페이지를 하나의 종점(terminal page)으로 정의하고, 시작 페이지로부터 각 페이지로의 최적의 경로를 찾아내면서 전체 site의 비용을 최소화할 수 있는 구조를 탐색하는 것이다. 또한 라그랑지안 릴렉세이션을 적용하여 경로 제약조건의 변화에 대해 효율적인 최적해의 변화를 검증하며, 웹의 구조적분석에 적합한지 여부를 적용하는 것이다. 본 연구에서는 웹에 대해 최적화 모델링을 입안 및 분석하였으며, 실험으로서 입안된 모델을 최적화 툴에 적용하여 최적구조화에 부합되는 결과를 얻을 수 있음을 입증하였다.

1. 서론

World Wide Web(웹)은 정보의 저장, 검색 및 습득에 있어 가장 중요한 매체로 자리 잡았다. 웹 검색엔진(Web search engine)은 일반적인 사용자들이 가장 많이 사용하는 정보 검색 및 습득 방법이다. 이러한 웹 검색엔진이 가진 여러 가지 문제점 중의 하나는 검색 결과를 통해 사용자가 해당 웹 페이지로 이동할 수 없는 경우가 발생한다는 것이며, 이러한 경우를 "Dead link" 혹은 “Invalid link"라고 한다[1]. 이러한 dead link에 대한 문제점은 이미 오래 전부터 언급되어 왔으며, 이는 비단 웹 검색엔진 만의 문제라고 말할 수 없다. Dead link가 일반적인 Web site의 신뢰도에 미치는 영향은 이미 Web search engine의 신뢰도에 미치는 영향과 동일한 수준에 이르고 있다[2]. 따라서 “접근성(Accessibility)” 혹은 “유용성(Availability)”은 웹 검색엔진 및 일반적인 웹 사이트의 품질을 평가하는 주요한 기준 중의 하나로 사용되기도 한다[3,4].

Dead link의 가장 중요한 원인으로는 정보의 휘발성을 들 수 있다. Web 페이지 상에서 하나의 정보가 생성되어서부터 소멸되기까지의 시간을 정보의 수명이라고 했을 경우, 그 평균 수명은 약 44일에 불과하다는 것은 이미 밝혀진바 있다 [5]. Koehler의 연구[6]에서 웹 사이트 내의 항목의 변화를 무시하더라도 웹 사이트의 평균 수명은 2년 정도임이 확인되었다. 즉, 지금 이 순간에도 Web 상의 정보는 생성되고 또 소멸되어가고 있기 때문에 지금 존재하는 정보를 내일은 찾을 수 없을 수도 있으며, 또 다른 정보로 대체될 수 있음을 의미한다. Web

server 상의 물리적인 문제는 일시적인 Web page의 접근 불가와 단편적인 문제를 발생시킬 뿐이지만, 정보의 휘발성은 dead link와 같은 문제의 직접적인 원인을 제공할 뿐만 아니라, WWW내에 존재하는 모든 정보에 대한 Web search engine의 coverage 성능이 40% 에도 채 미치지 못하게 되는 원인이 된다[7].

본 연구에서 웹을 보는 관점은 다음과 같이 formal 표현으로 정리할 수 있다. 전체 웹(WWW)은 유향 그래프 $G_w = (N_w, E_w)$로 인식되며, 웹 노드의 집합 $N_i = \{N_1, ..., N_n\}$ 및 웹 아크는 아크 함수 $x_{ij} : N^k \rightarrow \{0, 1\}$로서, $\forall (i, j) \in E_w$ 이는 유한 노드집합 N_w, 및 유한 아크 집합 E_w 의 순서쌍으로 인식하여 (i, j)로 표현할 수 있고 이때 i, j = {0, 1, 2, 3, ... , n-1}이며 n 은 유한갯수 $|N_w|$이다. 이때 웹 페이지와 웹 아크는 Uniform Resource Identifiers에 대응된다 [3, 10].

하나의 Web site는 node와 arc로 이루어진 directed graph로 생각할 수 있다. 이 때, node는 정보를 담고 있는 Web page로, arc는 Web page 간의 hyperlink로 생각할 수 있다. 일반적으로 하나의 Web site는 index 페이지를 가지고 있으며, 이 페이지를 통해 site 내 다른 Web page로 이동할 수 있게 된다. Index page로부터 특정 Web page로의 이동 경로를 하나로 제한할 경우, Web site는 하나의 Tree 구조로 인식될 수 있으며, 이 때 index page는 root node가 된다.

Index page를 제외한 나머지 Web page는 특정 정보를 저장하는 역할과 다른 페이지로의 이동경로의 역할을 담당한다. 각 페이지가 담고 있는 정보는 그 페이지의 중요도를 나타낼 수 있으며 중요도가 높은 페이지일수록 사용자가 접근하기 쉽도록 index 페이지에 가까운 곳에 위치시키는 것은 당연한 것이다. 예를 들어, 한 기업을 홍보를 담당하는 페이지는 소비자의 해당 기업에 대한 이미지를 결정할 수 있기 때문에 중요도가 높은 Web page라고 할 수 있다. 이러한 페이지를 index 페이지에서 직접 접근할 수 있도록 하는 것은 실제 여러 기업의 Web site에서 쉽게 확인할 수 있다. 그러나 이러한 Web page로의 접근이 불가능해진다면 해당 기업은 상당한 수준의 비용이 발생하게 된다. 따라서 이러한 문제는 Network Optimization Model중에서 Minimum Cost Network Flow Problem으로 모형화 할 수 있다. Minimum Cost Network Flow Problem은 선형계획 모델로 쉽게 표현할 수 있으며, 그 해 역시 Flow Conservation 조건을 만족할 경우 항상 존재한다[8].

본 논문의 주요 목적은 네트워크 모델과 정수 계획법을 이용하여 웹 사이트 구조를 최적화하되, "Dead Link"로부터 발생하는 비용을 최소화하고, 각 Web page의 중요도에 따라서 index page로부터의 거리를 달리함으로써 중요 웹 페이지에 대한 접근성(accessibility)을 높일 수 있는 조건을 적용하는 것이다.

논문의 구성은 다음과 같다. 2절에서는 관련 연구에 대해 간략하게 살펴보고, 3절에서는 Web site 구조화의 문제를 네트워크 모델과 정수계획법을 이용하여 formulation하는 과정에 대해 설명한다. 4절에서는 Lagrangian Relaxation을 이용하여 제약조건을 완화시키는 방법을 보여주고, 5절에서는 간단한 예제를 통해 문제를 formulation하는 과정을 보여주고 그 해가 기대비용을 낮춰준다는 것을 보여준다. 마지막으로 6절에서 논문의 간략한 요약과 함께 한계점을 지적하고 앞으로의 연구방향에 대해 언급하는 것으로 한다.

2. 관련연구

Web을 구조화하고자 하는 노력은 이미 오래 전부터 있어 왔다. Web catalogue 혹은 Superbook과 같은 계층적 구조화가 그 대표적인 예이다. Chen(1997)의 연구에서는 Generalized Similarity Analysis를 통하여 Web 상의 정보를 구조화하고 이를 시각화하여 사용자의 이해를 돕기 위한 시스템을 구축하였다[9]. MosaicG(1997) 프로젝트에서는 사용자가 방문한 Web page를 Tree 구조의 그래프로 나타내어 사용자가 방문했던 Web page로 쉽게 되돌아 갈 수 있도록 하였다[10]. 그러나 Chen의 연구는 Web site의 구조를 최적화하는 것이 목적이 아니라 Web 상에 존재하는 정보나 혹은 그 정보를 담고 있는 Resource들을 표현하기 위함이었으며, MosaicG 역시 단순히 사용자가 방문했던 페이지를 다시 쉽게 찾아갈 수 있도록 하기 위한 Web map의 역할일 뿐 Web page에 대한 Accessiblity를 높이기 위한 근본적인 해결책을 제시하지는 못했다. Botafodo, Rivlin과 Shneiderman의 연구(1992)에서는 Web site를 설계할 당시 설계자가 의도했던 계층 구조를 찾아내고, hypertext 구조의 또 다른 특징을 설계자에게 보여 줌으로써, user interface와 웹 사이트 구조를 개선하고자 하였다[11]. 그러나 이 방법은 이미 설계된 Web site의 구조를 개선할 수는 있으나, Web site 설계 시에는 어떠한 방안을 제시하지 못한다는 한계가 존재한다. 내용기반 구조 최적화에 관한 우리의 선행연구[14]는 거리평가척도(similarity measure)의 다양한 확장에 대해 실험하지 않았고, 개별링크나 웹 페이지의 깊이에 대한 제약조건을 변화하는 시도는 하지 않았다. 이런 점에서 Hop Constrained Min-Sum Arborescence 접근법은 시사하는 바가 크다. 원래 이 문제는 네트워크 설계와 Routing, Scheduling과 같은 분야에서 자주 다루어지는 문제이다로서 본 논문에서 중점적으로 다루고자 하는 문제 역시 HCMA 문제의 일환으로 해석될 수 있다. 지금까지 HCMA 문제를 해결하기 위한 Algorithm은 다양한 시도가 있었으나 웹을 대상으로한 연구는 없었다. 기술적 관점에서 Gouveia는 HCMA 문제를 정수계획법 문제로 모델링하고 이를 해결하기 위해서 Lagrangian 기반의 발견적 방법을 제안하였다[13]. Kawatra(2003)의 연구에서는 HCMA(Hop Constrained Min-Sum Arborescence) 문제를 풀기 위해 Lagrangian Relaxation과

Sub-gradient Optimization, Branch Exchange Heuristic을 이용하는 방법을 제안하였다[12]. 그러나 그 방법은 완전히 처음 설계하는 대상에는 적합하지만 링크나 페이지 구조의 제약조건이 부여된 웹 환경에 적용되기에는 무리가 있다.

3. 네트웍 모델링

본 논문의 목적은 우선 시작페이지(default page)를 제외한 모든 페이지를 하나의 종점 페이지(terminal page)로 정의하고, 전체 웹 사이트의 비용을 최소화할 수 있도록 시작페이지로부터 각 개별 웹페이지로의 최적의 경로를 찾아내는 것이다.

먼저, 다음 세 가지 조건을 모두 만족하는 Web Page를 종점 페이지(Terminal Page)로 정의한다.

(a) incoming link를 하나만 가진다.
(b) index page로부터의 unique한 path가 존재한다.
(c) index page로부터 terminal page까지의 link 개수는 미리 지정된 수 이하이다.

Notations

N : the set of terminal page t, $N=\{t \mid t=2, 3, \ldots, n\}$

l_{ij} : link from i^{th} page to j^{th} page

C_{ij} : establishment and maintaining cost of l_{ij}

D^t : outage cost of terminal page t

Q : link failure rate

d^t : depth constraints from index page to terminal page t

집합 N은 terminal page의 집합이며, C_{ij}는 l_{ij}를 설치할 때 발생하는 비용이다. D^t는 terminal page t로의 접근이 불가능할 때 발생하는 경제적인 손실을 의미한다. Q는 링크 l_{ij}를 통해 page i로부터 page j로 접근할 수 없을 확률을 의미하며, d^t는 index page로부터 terminal page t까지의 최대거리를 의미한다. 중요도가 높은 페이지일수록 D^t는 큰 값을 가지며, d^t는 작은 값을 가지게 된다.

Decision Variables

$Y_{ij}^t \in \{0,1\}$: if there is a direct link between page i and page j in the path from the index page to the terminal page t, Y_{ij}^t equals 1. Otherwise zero.

$X_{ij} \in \{0,1\}$: if there is a link between page i and page j in the optimal solution, X_{ij} equals 1. Otherwise zero.

결정변수 Y_{ij}^t는 index page로부터 terminal page t까지의 유일한 경로 위에 l_{ij}의 존재 여부를 나타내는 것으로, 경로 내에 l_{ij}가 존재할 경우 1, 그렇지 않을 경우 0의 값을 가진다. X_{ij}는 최적해 내에 l_{ij}가 존재할 경우 1, 그렇지 않을 때, 0의 값을 가지게 된다.

위의 notation들과 결정변수를 이용하여 다음과 같이 문제를 모델링할 수 있다.

$$Z_{IP} = \min \left\{ \sum_{i=1}^{n} \sum_{j=2}^{n} C_{ij} X_{ij} + \sum_{t=2}^{n} Q * D^t \sum_{i=1}^{n} \sum_{j=2}^{n} \right.$$

subject to

$$\sum_{i,j \in N} x_{ij} = 0 \quad \text{for} j = 0 \ \forall i \tag{1}$$

$$\sum_{i,j \in N} x_{ij} = 1 \quad if\ j \neq 0 \ \forall i \tag{2}$$

$$x_{ii} = 0 \quad \forall i \in N \tag{3}$$

$$x_{ij} + x_{ji} \leq 1 \quad \forall i, j \tag{4}$$

$$x_{ij_1} + \sum_{k=1}^{m-1} x_{j_k j_{k+1}} + x_{j_m i} \leq m-1, for\ 2 \leq m \leq |N| \tag{5}$$

$$\sum_{i=1}^{N} \sum_{j=2}^{N} Y_{ij}^t \leq d^t \quad for\ all\ t \in N \tag{6}$$

$$Y_{ij}^t \leq X_{ij} \tag{7}$$

$$X_{ij}^t \in \{0,1\} \quad for\ all\ i \in N \cup [1],\ and\ j \in N$$
$$Y_{ij}^t \in \{0,1\} \quad for\ all\ i \in N \cup [1],\ and\ j \in N,\ t \in N \tag{8}$$

목적함수 Z_{IP}는 링크의 설치비용과 링크의 파손으로 인해 발생하는 outage cost를 최소화하는 식이다.

제약조건 (1)은 루트노드로의 입력은 없어야 한다는 것이다. 즉, 누트노드의 부모를 용납하지 않는다. 제약조건 (2)는 트리를 위한 조건식이다. 즉, 부모노드가 여러 개가 있을 때 그중 하나만 선택되어야 트리가 만들어 진다는 것이다. 제약조건 (3)~(5)는 cycle 제거조건이다. 제약조건 (3)은 자기자신을 참조하는 self-cycle을 제거한다는 것이고, 제약조건 (4)는 인접한 2개의 노드가 상호교차 참조하는 경우를 배제하는 것이고, 제약조건 (5)는 3이상의 경로를 거쳐서 생기는 사이클을 제거한다는 것이

다. 제약조건 (6)은 index page로부터 terminal page까지의 링크 개수가 미리 정의된 수(d^t) 이하이어야 함을 나타낸다. Terminal page의 중요도에 따라 이 값은 미리 결정되어야 한다. 제약 조건 (7)은 optimal solution 내에 링크 I_{ij}가 존재할 때만 index page로부터 terminal page t로의 경로에 I_{ij}가 존재할 수 있음을 의미한다. 제약조건 (8)은 정수계획법이라는 의미이다.

4. Lagrangian Relaxation

본 연구의 또 다른 접근법으로서 Lagrangian Relaxation의 모델은 다음과 같이 구해졌다.

식 (2)의 relaxation : $\Theta(\sum_{i=1}^{N}\sum_{j=2}^{N} Y_{ij}^{t} - d_t)$

$$L(\Theta) = \min \left[\sum_{i=1}^{N}\sum_{j=2}^{N} C_{ij} + \sum_{t=2}^{N} Q * D_t \sum_{i=1}^{N}\sum_{j=2}^{N} Y_{ij}^{t} + \Theta(\sum_{i=1}^{N}\sum_{j=2}^{N} Y_{ij}^{t} - d_t) \right]$$

위의 $L(\Theta)$를 정리하면 다음 식이 된다.

$$L(\theta) = \min\left[\sum_{i=1}^{N}\sum_{j=2}^{N} C_{ij} + \sum_{t=2}^{N}\left(\sum_{i=1}^{N}\sum_{j=2}^{N}(\theta_t + Q^* D_t) Y_{ij}^{t} - \theta_t d_t\right)\right]$$

식 $L(\Theta)$중 다음 식 $m_t(\Theta)$를 제외한 모든 항들은 상수값이므로 모든 terminal page에 대한 $m_t(\Theta)$값을 최소로하는 Y_{ij}^{t}와 Θ_t의 값을 결정하면 $L(\Theta)$의 의 해를 찾게 된다.

$$m_t(\Theta) = \left[\sum_{i=1}^{N}\sum_{j=2}^{N}(\Theta_t + Q * D_t) Y_{ij}^{t}\right]$$

다음 식 $M_t(\Theta)$는 하나의 상품을 하나의 목적지에 최단거리 혹은 최소 비용으로 배달하는 single commodity flow problem 으로 해석될 수 있다. 따라서 index page로부터 각 terminal page로의 shortest path를 찾아내는 것과 동일한 문제이다. shortest path를 찾는 algorithm은 Dijkstra algorithm을 이용하며 node 간의 cost는 $(\Theta_t + Q * D_t)$를 이용한다.

$$M_t(\Theta) = \min\left[\sum_{i=1}^{N}\sum_{j=2}^{N}(\Theta_t + Q * D_t) Y_{ij}^{t}\right]$$

Θ값의 결정은 라그랑지안 승수 값에 대한 변화과정을 탐색하는 것이며, 이를 위해서는 앞서 설명한 subgradient optimization 방법을 이용한다.

5. 실험 예제 및 결과

실험에 사용한 웹 사이트 구조는 다음과 같다. 전체 7개의 노드로 구성되어있고, 트리형식으로 요약하면 다음 Fig. 7에 제시된 바와 동일하다.

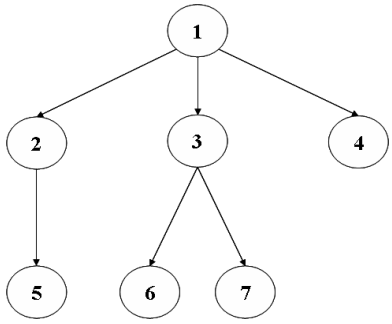

Fig 6. Illustrative Example Web Site Structure

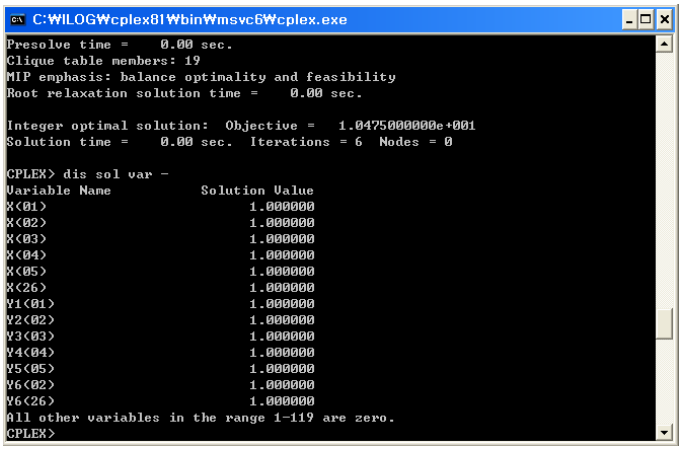

그림 7 C-Plex 실험결과

Link failure rate Q는 0.5로 하면 개별 웹 페이지에 이르는 비용은 다음과 같이 제시될 수 있다.

$$C_{ij} = \begin{pmatrix} 0 & 0.5 & 0.6 & 0.6 & 0.8 & 0.8 & 0.3 \\ 1.0 & 0 & 0.8 & 0.3 & 0.1 & 0.8 & 0.1 \\ 0.7 & 0.4 & 0 & 0.3 & 0.5 & 0.5 & 0.5 \\ 0.6 & 0.5 & 0.6 & 0 & 0.5 & 0.2 & 0.7 \\ 0.1 & 0.8 & 0.4 & 0.8 & 0 & 0.3 & 0.5 \\ 1.0 & 0.3 & 0.9 & 0.8 & 0.7 & 0 & 0.1 \\ 1.0 & 0.8 & 0.7 & 0.7 & 0.3 & 0.2 & 0 \end{pmatrix}$$

$$D^t = \begin{pmatrix} 2 \\ 6 \\ 9 \\ 10 \\ 5 \\ 8 \end{pmatrix} \qquad d^t = \begin{pmatrix} 4 \\ 3 \\ 2 \\ 2 \\ 3 \\ 2 \end{pmatrix}$$

Current Expected Cost = 34.3

최적화 모형에 대해 노드를 이상과 같은 모수값들을 가지고 최적화 도구(C-Plex)를 사용하여 얻은 결과는 다음과 같다. 그림 6의 내용 참조. 각 주어진 변수 X(ij)에 대하여 개별 종점페이지 (t)에 대한 결정변수 Yt(ij)는 for t= 0, 1, ..., 6, 다음과 같은 결과를 보여주었다. 즉, 최적 노드구조는 루트노드로부터 각각 N1, N2, N3, N4, N5 및 N2로부터 N6의 최적해가 얻어졌다. 그리고 결정변수에 따른 최장경로는 루트노드에서 N2, N2에서 N6이다. 실험에서는 데이터 량의 폭주로 인해 라그랑지안 릴렉세이션에 관한 실험은 수행하지 않았고, 단 경로수 감소에 따른 대안 최적해의 탐색과정을 확인하였다.

6. 결론 및 추후연구방향

웹 검색엔진이 가진 여러 가지 문제점 중의 하나는 검색 결과를 통해 사용자가 해당 웹 페이지로 이동할 수 없는 "Dead link" 혹은 “Invalid link"문제에 대응할 수 있는 최적화 모형과 그 변화에 대한 Lagrangian Relexation 모델링 과정에 대해 살펴보았다. 실험에서도 Hop 제약조건을 웹 페이지의 클릭 수로 즉, 루트 노드로부터의 깊이로 인식하였으며, 이 모형에서는 루트노드로 부터의 깊이가 줄어드는 조건의 제약식을 제시함에 따라 최적해의 변화와 대안 경로를 정확히 찾아내는 것을 확인할 수 있었다.

추후 연구과제로는 주어진 경로 중에서 제약조건의 개별 웹 페이지 별로 다양화하는 노력을 통해 웹 사이트 개인화 (personalization)에 대응하는 부분을 확인하고자 한다. 라그랑지안 릴렉세이션 모형에 대한 실험을 좀 더 세밀히 진행하고자 하며, 이를 통해 대안 노드를 선택하는 과정을 입증해야할 필요가 있다. 또한 실제 웹 사이트와 같이 큰 문제를 동일한 모형으로 해석을 시도하는 프로그램의 개발이 진행 중이다.

Acknowledgements

This work was partially supported by the Korea Science and Engineering Foundation (KOSEF) through the Advanced Information Technology Research Center (AITrc).

References

[1] Greg R. Notess, "On The Net: Dead Search Engines," Online, Vol. 26, No. 3, pp. 62-64, May/June 2002.
[2] Greg R. Notess, Search Engine Statistics: Dead Links Report, "http://www.notess.com/search/stats/dead.shtml", 2000.
[3] Hope N. Tillman, Babson Park, Evaluating Quality on the Net, “http://www.hopetillman.com/findqual.html”, 2003.
[4] Alastair G. Smith, Testing the Surf: Criteria for Evaluating Internet Information Resources. The Public-Access Computer Systems Review 8, No.3, 1997.
[5] Michael Lesk, Mad Library Disease: Holes in the Stacks, “http://www.lesk.com/mlesk/ucla/ucla.html”, 1996.
[6] Wallace Koehler, Digital libraries and World Wide Web sites and page persistence, Information Research, Vol. 4, No. 4, 1999.
[7] S. Lawrence and C. L. Giles. Accessibility of information on the web. Nature, 400 (July 8), pp.107~109, 1999.
[8] James R. Evans, Edward Minieka, Optimization Algorithm for Networks and Graphs, Marcel Dekker, Inc., pp.130, 1978.
[9] Chen, C. Structuring and Visualizing the WWW by Generalized Similarity Analysis. ACM Conference on Hypertext (Hypertext’97). Southampton, UK. ACM Press. pp. 177~186, 1997.
[10] Henzinger, M. R., Heydon, A., Mitzenmacher, M. and Najork, M., "On Near-uniform URL Sampling", Computer Networks, Vol.33, No.1 (2000), pp.295-308.
[11] Rodrigo A. Botafogo, Ehud. Rivlin, and Ben. Shneiderman. Structural analysis of Hypertexts: Identifying Hierarchies and Useful -Metrics. ACM Transactions on Information Systems, Vol.10, No.2, pp.142~180, 1992
[12] Rakesh Kawatra, A Hop Constrained Min-Sum Arborescence with Outage Costs, Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03), IEEE Computer Society, 2003.
[13] Luis Gouveia, Multi-commodity Flow Models For Spanning Trees with Hop Constraints, European Journal of Operational Research, Vol.95, pp.178~190, 1996.
[14] 이우기, 김승, 김한도, 강석호, “웰드와이드웹의 내용기반 구조최적화,” 한국경영과학회지, 제30권 제1호, pp. 187-198, 2005.