활동성, 신뢰성 기반의 Influence 지수 산정 알고리즘 설계

최창현, 박건우, 이상훈
국방대학교 전산정보학과
e-mail : budlelove01@hanmail.net

A Design of the Influence Value Computation Algorithm Based on Activity and Trust

Chang-Hyun Choi, Gun-Woo Park, Sang-Hoon Lee
Dept of Computer Science & Information, Korea National Defense University

요 약

집단지성을 이용한 지식검색 서비스는 개방적 구조와, 축적된 자료를 공유할 수 있다는 커뮤니티적인 특성으로 큰 인기를 얻고 있다. 하지만 방대한 지식공유속에서 사용자가 진정으로 원하는 답변 획득은 점점 더 어려워지고 있다. 최근 알고리즘적으로 가장 정교하다고 평가 받는 구글을 통해 상위에 랭크된 검색결과들 중에는 집단지성을 통해 구축된 위키피디아, Yahoo Q/A 과 같은 Social 검색엔진의 검색결과들이 상당수 존재한다. 본 논문은 대부분의 질문은 인간으로부터 문제해결의 실마리를 얻을 수 있다는 점과 온라인상의 사용자에 대한 연구를 통해 지식검색 서비스 사용자중 Influence를 찾는것에 목적이 있다. 이에 국내 Social 검색 엔진의 대표인 네이버 지식iN을 중심으로 지식검색내의 사용자 활동성과 신뢰성을 분석하고, 이를 기반으로한 Influence 지수 산정 알고리즘을 제안한다. 제안된 알고리즘을 통한 Influence 지수는 지식검색 서비스에서 문제 해결의 실마리를 가진 사용자를 찾는 중요한 지표가 될 것이다.

1. 서론

2002년 10월 네이버의 지식iN을 시작으로한 지식검색 서비스는 이후 엠파스, 야후 등과 같은 포탈들의 참여로 국내 검색 포탈들의 대표적 서비스로 성장하였다. 이러한 인기는 누구나 어떠한 주제에 대해서도 질문과 답변을 할 수 있다는 개방적 구조와, 이렇게 축적된 자료를 공유할 수 있다는 커뮤니티적인 특성에 기인한다. 하지만 방대한 자료의 구축은 사용자가 진정으로 원하는 답변 획득을 점점 더 어렵게 만들고 있다. 이러한 문제를 해결하기 위해 지식 검색 서비스의 결과물로 대변되는 답변 문서의 특성을 텍스트/비텍스트 관점으로 접근하는 연구[1],[2]가 수행되어 왔으나 결과물의 성격이 문서에서 동영상, 그림, 음성 등으로 다양해짐에 따라 문제 해결의 어려움을 겪고 있다. 본 논문은 지식 검색 서비스의 이러한 문제점을 집단지성의 특성과 Socoal Network 관점에서 풀어보고자 한다. 이는 공증된 객관적 수치를 바탕으로 한 높은 Influence 지수를 갖는 사용자라면 어떠한 형태(문서, 동영상, 음성 등)든 질 높은 질문/답변을 통해 지식공유의 근본적 목적에 부합될 수 있을 것이라는 가정에서 시작되며, 이를 위해 대표적인 국내 지식검색 서비스인 네이버 지식iN에서의 사용자 활동성, 신뢰성을 바탕으로 카테고리별 Influence 지수 산정 알고리즘을 제안한다. 산출된 Influence 지수는 문제 해결의 실마리를 가진 사용자를 찾음에 있어 훌륭한 지표가 될 것이다. 논문은 2장에서는 관련연구, 3장에서는 Social Network 기반 Influence 요소, 속성 추출 및 이를 통한 Influence 지수 산정 알고리즘을 제안한다. 4장에서는 실험을 위한 데이터 세트와 알고리즘의 일반화 가능성을 결과로 제시하고 마지막으로 결론 및 향후 연구 과제를 제시한다.

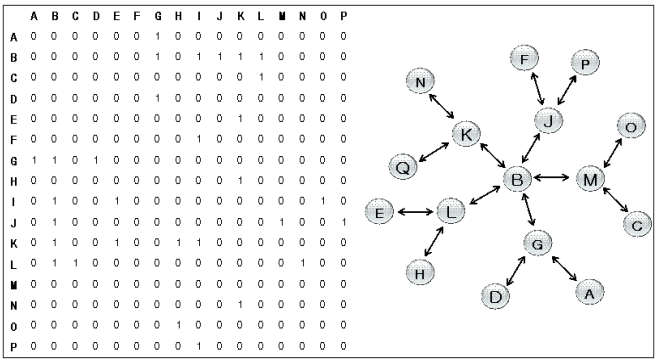
2. 관련연구

2.1 집단지성과 지식검색

지식과 정보의 구분이 모호해 지면서 등장한 지식의 여러 유형가운데 사회적 형식지는 온라인이란 환경에 힘입어 큰 영향력을 미치고 있다. 이러한 지식은 광의의 개념으로서의 일상생활과 관련된 지식, 다양한 지식생산자가 제공하는 지식, 상대적으로 불안정하고 유동적인 지식, 집합적으로 구성되는 지식이다[3]. 즉, 온라인에서의 지식은 우리의 일상생활과 관련된 정보, 상식, 조언까지도 포함하는 보다 확장된 개념으로서, 네티즌을 포함한 다양한 지식생산자들이 직접 제공하는 상대적으로 불안정하고 유동적인 지식이다. 또한 이를 기반으로 한 지식검색 시스템에서 제공하는 지식은 현재 나의 목적에 어떠한 의미가 있는가에 따라 현재 시점에서 창출되는 지식으로 인터넷의 장속에서 집합적으로 공유되고 끊임없이 구성되는 특성을 지닌다. 이러한 점은 집단지성의 발현과 관련이 깊다. 집단지성은 다수의 사용자가 개개인의 작업 및 지식을 공유하고 취합하여 일반적 사실을 도출해 낸다는 원리를 가지고 있다. 이 원리는 다수 사용자의 참여에 의해 어떤 사실에 대한 해결의 실마리를 얻는다는 것이 그 핵심이다.

2.2 Social Network 분석

Social Network는 최근 온라인을 중심으로 하여 사용자간의 관계를 맺고 있는 구조나 인간관계망으로 소개되고 있다. 즉, 사용자간의 연결이 존재한다는 것이며, 사용자는 적어도 한 가지 이상의 목적을 가지고 이를 이용한다. 이는 하나 이상의 상호의존적인 관계에 의해 구성된 개인 또는 집단으로 사회적 구조체로 정의된다. SNA(Social Network Analysis)는 Social Network의 형태와 특성을 알고리즘 적으로 연구하는 것으로 전체 관계망에서의 위치와 그 효과를 측정하는 위치적 접근법(positional approach)과 연결망의 직접적인 관계에 초점을 둔 관계적 접근법(relational approach)으로 분류된다[4]. 위치적 접근법은 사람들과의 사회적 관계에서 각자가 차지하는 위치 하나 하나를 가리켜 사회적 지위라고 부르며, 각각의 사회적 지위에 따라 기대되는 행위를 가리켜 사회적 역할이라고 한다. 관계적 접근법은 연결망 내 구성원들의 상호작용에 의한 전염효과, 즉 직접적인 관계 유무에 초점을 두어 '결속접근' 이라고 부르기도 한다. 분석방법은 노드간의 관계구조를 찾아내고 분석하기 위해 그래프 이론을 이용한 Sociometry를 사용하거나 수학적 방법을 이용한 계량적 분석을 사용한다. 수학적 방법의 기본은 행렬과 그래프의 이해이다. 구성원 (i, j) 사이의 관계가 있고 없음을 1과 0으로 나타내는 행렬을 인접 행렬(adjacency matrix)이라고 부르며 완전 연결망의 기본 형태이다. 행렬의 항(cell)은 i 로부터 j 에에 이르는 관계를 표현한다.



(그림 1) 인접행렬과 그래프

(그림 1)의 예에서 A, G항이 1인데, 이는 A로부터 G에 이르는 관계가 있다는 뜻으로(예: A가 G를 친구로 선택한 경우) 그래프에서는 A와 G간 연결된 화살표로 표현된다. 인접행렬의 대각선은 자기 선택 유무를 나타내는 것으로 일반적으로 0으로 처리 한다. 한 노드가 다른 노드와 연결될 때 연결된 노드의 수를 연결정도(degree)라 한다. 한 노드에서 다른 노드로 나가는 수의 합을 외향 연결정도(out-degree)라하고 다른 노드로부터 들어오는 수의 합을 (in-degree)라 한다. 전체 연결망에서 특정한 노드의 내향 연결정도가 큰 경우 이 점은 중요한 역할을 하는 경우가 많다. (그림 1)의 예에서 B의 외향 연결정도는 '5'이며, 내향 연결정도도 '5'이다[5].

2.2.1 중앙성

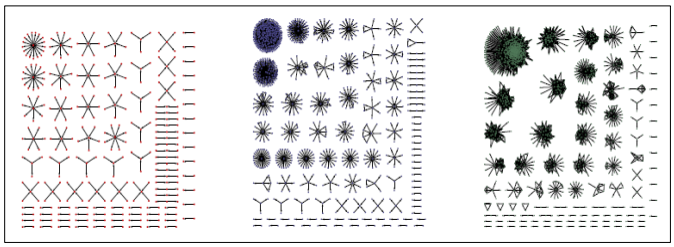
연결망 기법은 결속, 중앙성, 구조적 등위성 등의 주제로 구별할 수 있다. 이중 중앙성(centrality)이란 한 노드가 맺은 관계의 정도를 통해 권력과 영향력이라는 개념과 연결되어 가장 많이 쓰이는 지표 가운데 하나이다. 프리만(Freeman)은 중앙성을 지역 중앙성(local centrality)과 전체 중앙성(global centrality)으로 구분하였다[6]. 한 노드가 지역적으로 중앙성을 갖는다는 것은 그 노드가 속한 관계망 환경에서 다수의 다른 노드와 연결 관계를 갖는 것, 즉 그 노드가 다수의 이웃 노드를 직접적인 연결로서 가지고 있다는 것을 의미한다. 반면 한 노드의 전체 중앙성은 그 노드가 전체적인 관계망 구조에서 전략적으로 중요한 위치를 가진다는 것을 뜻한다[7]. 중앙성은 관계의 연결 정도를 통해서 계산되는 수치로 내향 연결정도를 이용한 내향 중앙성과 외향 연결정도를 이용한 외향 중앙성의 합에 전체 연결정도의 합을 통해 얻은 중앙성을 나누어서 구한다[5].

$$C_i = \sum_{j=1}^{n}(Z_{ij} + Z_{ji}) / \sum_{i=1}^{n}\sum_{j=1}^{n}(Z_{ij}) \ , \ 0 \le C_i \le 1 \quad (1)$$

3. Influence 지수 산정

3.1 Influence 요소와 속성 산정

Influence의 사전적 의미는 '어떠한 사물의 효과나 작용이 다른 것에 미치는 힘'이다. 크리스 와이드너는 온라인에서의 영향력을 가상의 공간에서 구성원의 잠재력을 끌어내고 긍정적인 변화의 기운을 만들어 내는 사용자의 능력이라 하였다[8]. 다양한 목적으로 지식 검색 서비스를 이용하는 사용자는 집단지성의 발현을 통한 문제 해결의 실마리를 찾고자 한다. 이것은 또한 지식 검색 서비스의 최대 목적이 된다. 국외 대표적 지식공유 검색 사이트인 Yahoo의 질문/답변에 대한 특성 분석[4]은 흥미 있는 결과를 제시한다. 총 25개의 카테고리에 대한 질문/답변의 길이, 답변 실마리의 내용, 질문자/답변자 오버랩을 통한 k-means 클러스터링은 (그림 2)와 같은 3가지로 크게 분류된다.



(가) Factual (나) Advice (다) Forum

(그림 2) 지식 검색 Social Network구조

첫째는 'Forum'클러스터로 "스포츠에서 누가 이길 것 같은가?", "가장 감명 깊게 봤던 영화는 무엇인가?" 등이 이러한 클러스터에 해당된다. 두 번째는 조언을 찾고 제공하며, 일상적이면서도 다소 전문적일 수 있는 'Advice'클러스터로 패션, 결혼/이혼 상담 등이 이에 해당한다. 마지막으로

어떠한 사실 여부에 대한 질문/답변으로 과학적 사실, 프로그래밍, 학문 등이 해당되는 'Factual' 클러스터이다. 이러한 클러스터내 사용자들은 유사한 질문/답변 행동양식과 지식의 신뢰기준[9]을 가지고 있다. 또 다른 연구[10]에서는 가상공간내의 핵심 멤버 찾기를 사용자의 관심과 Tag 수로 보았다. 이들을 통해 본 논문에서는 Influence의 요소를 활동성과 신뢰성으로 정의한다. 또한 국내 지식검색 서비스로부터 공통적으로 추출 가능한 사용자의 질문수, 답변수, 답변채택수, 질문확정수를 속성으로 정의한다.

3.2 Influence 활동성

지식검색 서비스는 사용자들이 지식을 검색, 공유 및 활용하며, 직접 만들어가는 서비스이다. 이러한 지식검색 서비스의 시스템은 이용자가 궁금한 것을 '질문'하고, 질문기간 동안 다른 이용자가 이것에 대해 '답변'을 하는 형식으로 이루어진다. 질문기간이 지나면 수많은 답변들중 하나를 '답변채택'한다. 이것이 하나의 공유 가능한 지식으로 완결되어 저장 된다. 이렇게 저장된 지식들은 이용자들이 함께 공유하고, 검색하여 활용할 수 있다. 즉 네티즌들이 지식, 정보의 생산자와 소비자로 동시에 활동할 수 있는 공간이 된다. 이를 수학적으로 해석하면 다음과 같다[11]. 사용자들마다 알고 있는 지식을 Xk = a + nk 로 표현할 때 a는 옳은 지식이고 nk는 잘 모르는 지식을 의미한다. 여기서 k는 사용자의 번호를 의미한다. 지식 X에 대해 1부터 k까지의 사용자가 알고 있는 지식을 모두 합하면 식(2)와 같이 표현된다.

$$x_{CI} = \sum_{k=1}^{K} x_k = Ka + \sum_{k=1}^{K} n_k \qquad (2)$$

식 (2)에서 옳은 지식은 그 양이 증가할수록 선형적으로 증가하지만 잘 모르는 지식은 모두 같은 지식이 아니며 상호 간섭이 적기 때문에 서로 독립적이라고 볼 수 있다. 지식검색에서의 사용자 질문, 답변 행위도 서로 독립적이다. 사용자 i 의 활동성을 질문과 답변에 대한 회수로 계산시 아래와 같은 수식을 유도 가능하다.(3)

$$U_{i\_Activity} = \sum_{n=1}^{p} (u_{i\_question})_n + \sum_{m=1}^{q} (u_{i\_answer})_m \qquad (3)$$

(사용자 i = {1, 2, 3, ... , k} 일 때,

p, q 는 개인별 질문수, 응답수에 대한 변화값이며,

$u_{i\_question}$: 사용자 i의 질문, u_{i_answer}: 사용자 i의 답변이다.)

3.3 Influence 신뢰성

본 연구에서는 신뢰성에 대해 질문자의 답변자 채택을 중요 요소로 고려한다. 지식검색에서 지식은 관련 연구에서 언급하였듯이 사용자의 목적에 어떠한 의미가 있는가에 따라 현재 시점에서 창출되는 지식이다. 이러한 점을 고려시 질문에 대한 수많은 답변중 질문자가 답변을 선택하는 것은 질문자에게 있어 답변 채택자와 신뢰성이 있는 관계의 성립이라 볼 수 있다. 본 논문에서는 카테고리별 Influence의 질문/답변 행동양식과 신뢰성 기준의 상이함을 고려 카테고리별 질문자와 답변채택자와의 관계를 통한 Social Network를 구축한다. 이러한 결속망에서 중앙성이 있는 노드를 찾아낸다. 한 노드가 Network내에서 지역적으로 중앙성을 갖는다는 것은 그 점이 속한 환경에서 다수의 다른 노드와 연결 관계를 갖는다는 것이다. 이는 지식공유에 있어 사용자의 활동성과 신뢰성을 의미하는 것이다. 또한 노드를 향해 오는 내향 중앙성(In-Centrality)은 신뢰성의 요소로 작용되며, 밖으로 나가는 외향 중앙성(Out-Centrality)은 사용자의 지식공유 구축에 대한 활동성 요소로 작용된다. 내향 연결정도와 외향 연결정도는 수식(4)와 같이 계산된다[5].

$$indegree_{ik} = \sum_{j=1}^{N} Z_{ijk} = Z_{ik} \qquad (4)$$

$$outdegree_{ik} = \sum_{j=1}^{N} Z_{ijk} = Z_{ik}$$

Where, Z_{ijk} : k 연결망에서 i 사용자로부터 j 사용자와의 관계

$indegree$: 사용자 i가 다른 모든 사용자들 j로부터 받는 답변채택자 수

$outdegree$: 사용자 i로부터 다른 모든 사용자 j에게 가는 답변채택자 수

3.4 Influence 지수 산정 알고리즘

국내 네이버 지식iN은 질문/답변의 범주를 총 11개의 카테고리와 144개의 세부 카테고리로 분류한다. 각 카테고리에서의 질문/답변 행위는 [4]에서와 같이 카테고리별 상이한 활동성과 신뢰성 기준을 가진다. 이에 사용자 k 로 구성된 연결망내에서 카테고리별 Influence 활동성, 신뢰성을 기반으로한 사용자 i 의 Influence 지수 산정 알고리즘을 수식(5)와 같이 제안한다.

$$U_{i\_IV} = \alpha \sum_{n=1}^{p} (u_{i\_question})_n C_{i\_outdegree} + (1-\alpha) \sum_{m=1}^{q} (u_{i\_answer})_m C_{i\_indegree} \qquad (5)$$

(α : 카테고리별 질문/응답수의 특성을 고려한 가중치, $0 \le \alpha$

C : 카테고리내 사용자 Node의 중앙성 지수

$$C_{i\_indegree} = \frac{in\,\mathrm{degree}_{ik}}{k-1}, \quad C_{i\_outdegree} = \frac{out\,\mathrm{degree}_{ik}}{k-1}$$

k : 네트워크에 존재하는 총 노드수, $0 \le C_i \le 1$)

사용자의 활동성 증가는 질문/답변수의 증가를 가져온다. 이와는 상반되게 질문/답변채택수의 증가로 추가되는 Node는 전체 Social Network의 중앙성 지수 감쇄 효과를 가져온다. 가중치 산출은 인공신경망의 역전파 알고리즘을 이용한다. 이를 통해 카테고리별 질문/응답수의 특성이 가질 최적의

가중치를 산출한다. 인공신경망에선 은닉노드의 활성화 함수로서 시그모이드 함수를 이용하며, 수식(6)과 같다.

$$f(sum_j) = \frac{1}{1 + e^{-sum_j}} \quad (6)$$

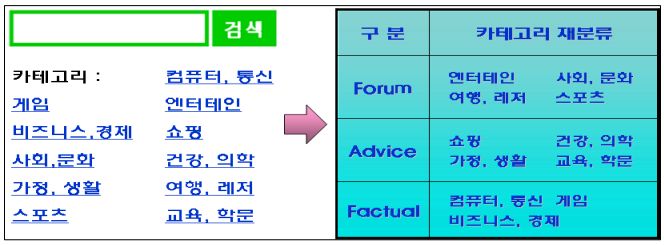
이와 같이 출력층의 신경세포들도 은닉층으로부터 입력을 받아 목표 값과 활성값의 차인 오차를 구한다. 이후 오차 신호를 계산하고 신경세포 i 와 j 간의 연결강도들(w_{ij})을 변화시키면서 카테고리별 최적의 가중치를 구한다.

4. 실험 및 결과

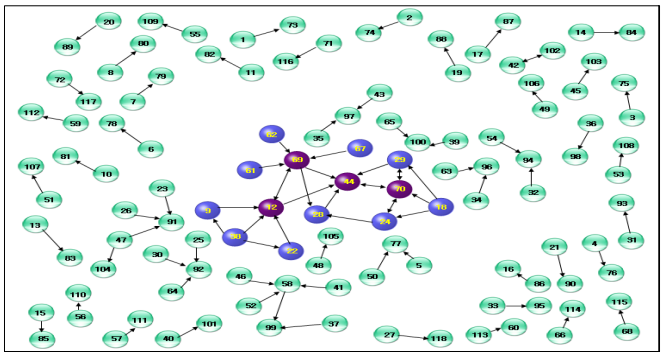
4.1 데이터 셋

제안한 알고리즘의 실험을 위해 네이버 지식iN의 카테고리를 <표 1>과 같이 Factual, Advice, Forum 으로 분류하였다.

<표 1> 사용자 활동성, 신뢰성에 따른 카테고리 재분류

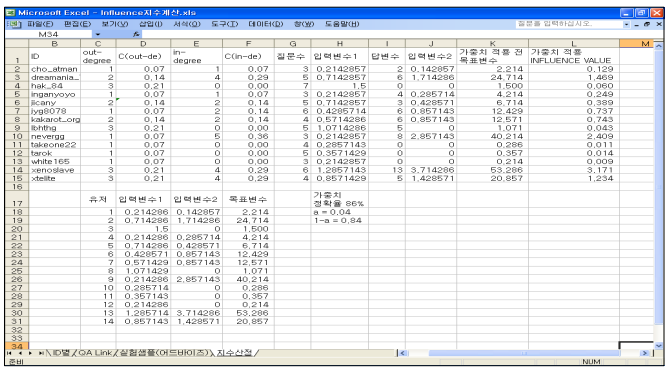


네이버 지식검색에서는 하루에만 수천-수만 개의 질문/답변이 이루어진다. 질문/답변을 통한 실험용 Social Network를 구축하기 위해 분류된 카테고리 내에서 카테고리 범주에 속하며, 적절한 답변들이 제시된 질문/답변 세트를 100개 수집하였다.



(그림 3) 지식검색 질문/답변자의 Social Network

<표 2> 가중치 및 산정된 Influence 지수



또한 알고리즘 적용에 대한 실험은 Factual 클러스터의 컴퓨터, 통신 카테고리내 세부 카테고리인 프로그래밍으로 한정하였다. (그림 3)은 질문/답변채택으로 구성되어진 176명 사용자의 Social Network이다.

4.2 Influence 지수 산정 결과

<표 2>는 구성된 결속망에 제안된 Influence 지수 산정 알고리즘을 적용한 결과를 나타낸다. '프로그래밍' 카테고리내 Influence 지수는 사용자 69, 44, 12, 70번의 순서로 상위 랭크 되었으며, 가중치 값은 질문 : 0.04, 답변 : 0.84를 나타낸다.

5. 결론 및 향후 연구

최근 인터넷 환경은 분산 컴퓨팅 및 서비스 기반 인터넷으로 빠르게 변화하고 있다. 이러한 변화를 가속화시키는 중요 기술로 집단지성은 다양한 분야에서 활용되고 있다. 집단지성을 이용한 지식검색 서비스는 다양한 사용자들의 지식 및 경험이 축적되어 대중적으로 활용 가능한 형태의 정보를 만들어 낼 수 있다는 것이 그 특징이다. 본 논문에서는 지식검색 사용자의 사용목적과 가장 부합된다고 할 수 있는 Influence 찾기를 사용자의 활동성과 신뢰성을 기반으로 Socail Network 분석기법을 적용하여 해결하고자 하였다. 본 논문에서 제안한 알고리즘을 통한 Influence 지수는 지식검색 서비스에서 문제 해결의 실마리를 가진 사용자를 찾는 지표가 된다. 또한 어떤 한 가지 이상의 목적을 가지고 Social Network를 이용하는 사용자의 행동양식을 분석 함에도 유용하게 적용 가능하다. 향후 데이터 세트를 추가하여 카테고리별 사용자 행동양식을 보다 세밀하게 분석하고, 하나의 지식으로 저장된 질문/답변 세트에 대한 다른 네티즌의 반응을 추가 요소로 반영한다면 지식검색 서비스의 만족도를 보다 향상시킬 수 있을 것으로 본다.

참고문헌

[1] 박소연, 이준호, 전지운 "지식검색 서비스 개선을 위한 문서의 적합도 및 신뢰도 분석" 한국문헌정보학회지 제40권 제2호
[2] Jiwoon Jeon, W.Bruce Croft, Joon Ho Lee, Soyeon Park "A Framework to Predict the Quality of Answers with Non-Textual Features" SIGIR 2006
[3] 김희연 "정보사회에서의 지식과 지식검색에 대한 고찰" 정보통신정책 제 18권 14호
[4] Lada A.Adamic, Jun Zhang, Eytan Bakshy, Mark S.Ackerman "Knowledge Sharing and Yahoo Answers:Everyone Knows Something" WWW 2008
[5] 김용학 "사회연결망 분석", 박영사, 2003
[6] Freeman L "Centrality in Social Networks: A Conceptual Classification", No.1, 1979
[7] John P, Scott "Social Network Analysis: A Handbook" , SAGE Publications, 2000
[8] 크리스 와이드너 "영향력", 리더스북, 2008
[9] Soojung Kim, Jung Sun Oh, Sanghee Oh "Best-Answer Selection Criteria in a Social Q&A site from the User-Oriented Relevance Perspective" ASIST 2007
[10] Haiqiang Chen, Xueqi Cheng, Yue Liu "Finding Core Members in Virtual Communities" www2008
[11] Szuba T "Computational Collective Intelligence", Wiley and Sons NY, 2001