Map/Reduce를 이용한 블로그 연결망 분석 시스템 설계

정회원 조 인 휘*, 준회원 박 재 균*

The Design of Blog Network Analysis System using Map/Reduce Programming Model

Inwhee Joe* *Regular Member*, Jae-Kyun Park* *Associate Member*

요 약

최근, 인터넷의 발달로 인해 온라인 사회연결망이 증가하고 있으며 이 중 블로그 서비스가 대표적이다. 본 논문에서는 블로그 연결망을 분석하기 위한 방법론을 제시하며, 대용량의 연결망 데이터를 안정적으로 분산 처리할 수 있는 방안을 제시한다. 우선, 각 연결망 데이터의 시간 경과에 따른 행위 가중치의 값을 보정하여, 최근의 행위가 과거의 행위보다 높은 연결강도를 가질 수 있도록 한다. 둘째로, 명시적으로 연결행위를 가지지 않은 블로그의 경우 블로그의 성격을 대표할 수 있는 키워드간의 유사도를 추출하여, 묵시적으로 연결망 내에 흡수하도록 한다. 따라서 이전의 방법론과는 달리 더 많은 블로그 노드 간의 연결을 분석할 수 있다. 본 논문이 제시한 블로그 연결망 분석 시스템의 설계로 기존에 제시되었던 방법론보다 약 40% 더 많은 블로그 간 연결망을 추출 할 수 있음을 보였으며, 시간의 흐름에 따른 연결강도의 변화에 대한 타당성을 입증하였다.

Key Words : Blog, Social Network Analysis, MapReduce

ABSTRACT

Recently, on-line social network has been increasing according to development of internet. The most representative service is blog. A Blog is a type of personal web site, usually maintained by an individual with regular entries of commentary. These blogs are related to each other, and it is called Blog Network in this paper. In a blog network, posts in a blog can be diffused to other blogs. Analyzing information diffusion in a blog world is a very useful research issue, which can be used for predicting information diffusion, abnormally detection, marketing, and revitalizing the blog world. Existing studies on network analysis have no consideration for the passage of time and these approaches can only measure network activity for a node by the number of direct connections that a given node has. As one solution, this paper suggests the new method of measuring the blog network activity using logistic curve model and Cosine-similarity in key words by the Map/Reduce programming model.

Ⅰ. 서 론

사회연결망(social network)은 임의의 사회 내에 존재하는 구성원들 간의 관계 정보를 이용하여 그 사회를 연결망(network)으로 표현한 것이다[1]. 이러한 사회적 관계를 연구하여 그 사회가 가지고 있는 고유의 특징을 도출하는 것을 사회연결망 분석(social network analysis) 이라고 한다. 과거에는 사회연결망 내의 구성원 사이의 관계 여부를 나타내는 데이터를 확보하는데 큰 어려움이 있었다. 따라서 기존 사회연결망 연구의 초점은 구성원들 간의 관계에 대한 구체적인 데이터 없이 사회연결망이 과연 어떠한 구조와 특징을 가지는

* 한양대학교 컴퓨터공학부 이동네트워크 연구실 (iwjoe@hanyang.ac.kr)
논문번호 : KICS2010-02-054, 접수일자 : 2010년 2월 3일, 최종논문접수일자 : 2010년 8월 31일

가를 모델링하는 데에 있었다.

인터넷의 발달로 인하여 온라인상에서도 이러한 사회연결망이 나타났다. 온라인 사회연결망(online social network)의 가장 큰 특징은 기존의 사회연결망과는 달리 구성원들 간의 관계를 설명할 수 있는 구체적인 정보와 그 관계에 의하여 주고받는 정보를 데이터베이스에 저장하고 있다는 것이다. 이러한 특징 때문에 최근 들어 온라인 사회연결망을 이용한 분석이 활발히 이루어지고 있다[2].

블로그(Blog)는 사용자가 자신의 글을 온라인상에 게시할 수 있는 개인 웹 사이트이며 블로그 서비스(blog service)는 사용자가 블로그를 생성하고 운영할 수 있도록 지원해 주는 서비스이다. 이러한 블로그 서비스는 블로그(blog)로 이루어진 대표적인 온라인 사회연결망의 하나로 최근 들어 다양한 비즈니스가 등장하고 있으며[9], 온라인 사회연결망 데이터들을 사용하여 사회 연결망의 위상 구조에 대하여 정밀한 분석을 위한 연구가 가능하게 되었다[3]. 블로그 서비스를 사용하는 사용자들은 서로 간 다양한 관계를 맺을 수 있으며, 이 결과 사회연결망이 형성된다[7]. 이러한 구성원 간의 관계의 정도에 대한 정보는 기존의 사회 연결망에서는 존재하지 않는 것으로, 이를 이용하여 사회연결망에 대한 기존의 연구에 비해서 실제적인 연구가 진행되고 있다[4].

본 논문에서는 구글(Google)이 발표한 Map/Reduce 프로그래밍 모델을 이용하여 대용량 분산처리 방식을 제안하고자 하며, 또한 대용량 기초 데이터의 왜곡현상을 해결하기 위해 시간가중치를 부여하여 블로그 간 연결망 데이터마이닝의 새로운 모델을 제시하고자 한다.

Ⅱ. 연결망 분석 시스템 설계

2.1 관계분석 연결 행위 정의 및 가중치 설계

블로그 연결망 내에서의 블로그 간 상호 연결 관계는 각각의 게시된 글에 대하여 (1) 조회하거나, (2) 댓글을 남기거나, (3) 스크랩을 하거나, (4) 엮인 글 달기와 같이 4가지의 행위로 이루어진다고 볼 수 있다[5,6]. 조회하기는 게시된 글을 읽는 행동이고, 댓글 남기기는 게시 글에 자신의 의견을 남기는 행동이고, 스크랩하기는 게시된 글을 자신의 블로그로 등록하는 행동이고, 엮인 글 달기는 연관된 새로운 내용을 자신의 블로그 내에 게시글로 작성하는 행동이다.

이러한 행위는 항상 같은 빈도로 발생하지 않기 때문에 연결 행위의 분석을 위해 각각의 개별적인 행위에 대한 가중치가 부여 되어야 한다, 또한 같은 행위라고 하더라도 주체, 피 주체와 같이 행위주체의 방향에 따라서 다른 가중치가 부여 될 수 있다. 이렇게 행위의 빈도와 행위 주체에 따른 가중치를 고려하여 기본적인 가중치 설계를 아래의 클래스 테이블을 마련 할 수 있다.

블로그 간 연결점수는 이러한 모든 연결 관계 행위 점수의 합으로 계산될 수 있다. 하지만 이러한 단순한 합산으로만 각각의 블로그 간 연결점수를 낼 경우 특정시점에 기록된 행위 점수의 합이 계속하여 증가하는 경향이 있다. 따라서 오래전에 일어난 연결 행위가 계속적으로 누적되었을 경우 이 점수의 합이 최근에 일어난 연결 행위의 점수의 합보다 높아짐으로써, 결국 최근 연결망분석에 대해서는 왜곡현상이 일어날 수 있다.

본 연구에서는 이러한 문제를 해결하기 위해, 각 행위에 대해서 로지스틱 성장곡선(Logistic Curve) 을 적용하여, 최근의 행위가 과거의 행위보다 높은 점수를 가지도록 하는 방법을 사용한다. 가중치가 특정 임계치 시간까지는 완만하게 감소하다가, 또 어느 정도의 시간 임계치에 다다르게 되면 급속하게 가중치가 감소되게 함으로써, 시간의 흐름에 따른 가중치 보정을 할 수 있다.

표 1에서 설계한 가중치 설계에 로지스틱 성장곡선을 적용할 경우 블로그 연결 행위 클래스에 접목할 경우 시간이 지남에 따라 이전 점수의 합산이 계속적으로 누적되어 점수의 왜곡이 생기는 문제를 보정할 수 있다. 표 2는 가중치 1차 설계에 로지스틱 성장곡선을 적용하여 구한 가중치 변화를 보여주고 있다.

앞서 설계된 표 1에 시간가중치가 적용된 최종 가중치 테이블은 아래와 같다.

표 1. 연결행위 정의 및 가중치 1차 설계

행위 설명	행위코드	주체 가중치	피주체 가중치
Ua 가 Ub 의 글 B1을 조회함.	RD	0.4	0.2
Ua 가 Ub 의 글 B1에 댓글을 달음.	CM	0.7	0.4
Ua 가 Ub 의 글 B1을 스크랩함	SC	0.6	0.3
Ua 가 Ub 의 글 B1에 엮인글을 생성함	TR	1	0.6

표 2. 로지스틱 성장곡선이 적용된 시간 가중치 변화 표

시간	0	2	4	6
가중치	1.00000	0.99679	0.99156	0.98310
시간	8	10	12	14
가중치	0.96956	0.94829	0.91585	0.86847

표 3. 시간가중치가 적용된 연결 관계 행위 가중치 테이블

행위 설명	행위코드	주체 최대 가중치	주체 최소 가중치	피주체 최대 가중치	피주체 최소 가중치
Ua 가 Ub의 글 A1을 조회하였다.	RD	0.4	0.2	0.2	0.05
Ua 가 Ub의 글 A1에 댓글을 달았다	CM	0.7	0.3	0.4	0.2
Ua 가 Ub의 글 A1을 스크랩하였다	SC	0.6	0.4	0.3	0.15
Ua 가 Ub의 글 A1에 엮인 글을 생성하였다	TR	1	0.5	0.6	0.3

2.2 블로그 사용자간 키워드 유사도 설계

앞서 설명한 연결망 분석방법은 직접적인 서로간의 행위에 의해서 일어나는 연결정보만 분석할 수 있다. 바꿔 말하자면 명시적으로 행위가 일어나지 않은 각각의 블로그는 연결망 분석 범주에서 제외될 수밖에 없고 분석이 불가능하다는 한계점을 가지고 있다.

본 논문은 이와 같은 문제를 해결하기 위해 명시적인 연결행위가 없더라도, 블로그 연결망에 암묵적으로 흡수되게 할 수 있도록, 유사도 검사를 사용하였다. 즉, 블로그 내에 있는 모든 게시물은 블로그의 성격을 나타낼 수 있는 대표적인 속성으로 볼 수 있으며, 이러한 속성은 게시물 태그(키워드)에 가장 강하게 나타나며, 각 블로그 간의 태그 유사도를 계산함으로써, 서로간의 관계 정도를 분석할 수 있다.

블로그 간 태그유사도를 구하기 위해서는 먼저 해당 블로그가 소유하고 있는 모든 태그를 구하고 각 태그별 사용 횟수를 측정하여야 하는데, 이것을 Term Frequency(TF)라 한다. 이후 각 블로그 사용자간 모든 TF 가 측정되면 Vector화 한 뒤 사용자별 코사인 유사도(Consine Similarity)를 이용하여 서로간의 태그 유사도 정도를 추출할 수 있다.

코사인 유사도 공식을 사용하여 아래와 같은 Term Frequency를 가지는 블로그 사용자간 키워드 유사도를 구할 수 있다.

표 4는 블로그 사용자 A와 블로그 사용자 B의 Term Frequency 정보를 Vector화하여 나타낸 도표이다. 사용자 A는 총 7번의 태그를 작성하였으며 키워드1은 1번, 키워드3은 4번, 키워드4는 2번 사용하였다. 마찬가지로 사용자 B는 총 5번의 태그를 작성하였으며, 키워드1은 사용하지 않았고, 키워드2는 2번, 키워드3은 3번 사용하였다.

표 4. 블로그 사용자별 Term Frequency Vector 예시

행위 설명	키워드1	키워드2	키워드3	키워드4
블로그 A	1/7	-	4/7	2/7
블로그 B	-	2/5	3/5	-

위와 같은 Term Frequency Vector를 코사인 유사도 수식을 이용하여 서로간의 태그유사도를 구할 경우 최소 0과 최대 1사이의 값인 0.726 값이 도출될 수 있다.

2.3 스키마 설계

블로그 연결망을 분석하기 위해서 본 논문은 행단위로 자료를 저장하는 RDMS가 아닌 로우 키를 기반으로 복수개의 컬럼 키와 값을 유동적으로 가질 수 있는 컬럼 기반 분산데이터 베이스 시스템인 HBase를 사용하도록 한다.

2.3.1 블로그 간 연결정보 저장 테이블

연결정보 저장을 담당하는 스키마는 주체자의 UID, 피 주체자의 UID, 행위코드, 행위시간 정보를 가질 수 있도록 설계하였으며, 행위시간정보의 경우 로지스틱 성장곡선 모델을 이용하여, 시간이 지남에 따른 가중치 조절 대상 데이터로 활용되어 분석한다.

표 5. 블로그 간 연결정보 저장 테이블 [USER_TRANSACTION])

Row Key		일련번호(Sequence)
Column Family:BASIC	Subject	주체 블로그 UID
	Object	피 주체 블로그 UID
	Verb	행위코드
	Starget	주체대상 Object ID
	Otarget	피 주체대상 Reference ID
Column Family:TIME	Action	행위시간 TimeStamp
	Process	처리시간 TimeStamp
Extension		기타 부가확장 요소 저장

2.3.2 블로그 별 키워드 통계저장 테이블

블로그 사용자별 키워드 유사도를 통한 연결망 추출을 위하여 각 블로그 사용자별 Term Frequency 테이블 설계가 필요하다. Column 기반의 데이터베이스 모델의 특징을 이용한 위의 스키마는 사용자별로 n 개의 키워드 컬럼이 생성될 수 있으며, 각 키워드별 사용횟수가 저장되고 사용자별 전체 키워드 사용 횟수가 저장되게 한다.

2.3.3 키워드별 블로그 사용자 저장 테이블

키워드 유사도의 경우 블로그 서로간의 중첩된 키워드가 1개라도 존재하지 않을 경우 유사도 계산의 범주에서 제외되어야 한다. 그렇지 않을 경우 각 사용자별로 모든 사용자와의 키워드 유사도 관계를 검사하게 되는데, 이는 상당히 비효율적인 분석방법이 될 수밖에 없다. 따라서 키워드별로 사용자를 다시 묶어서 실제 연결 관계가 성립될 수 있는 정보만을 추출하여 데이터베이스에 저장하도록 한다.

표 6. 블로그 사용자별 키워드 Term Frequency 테이블 [USER_KEYWORDS]

Row Key		블로그 UID
Column Family: KEYWORDS	[KEYWORD1]	[KEYWORD1]의 Frequency
	[KEYWORD2]	[KEYWORD2]의 Frequency
	…	…
	[KEYWORDn]	[KEYWORDn]의 Frequency
Column Family: SUMMARY	COUNTER	Total Frequency

표 7. 키워드별 블로그 사용자 저장 테이블 [KEYWORD_USERS])

Row Key		키워드
Column Family:USERS	[UID1]	블로그 UID1
	[UID2]	블로그 UID2
	…	…
	[UIDn]	블로그 UIDn

2.3.4 블로그 사용자간 키워드 유사도 결과 저장 테이블

마지막 테이블은 사용자간 키워드 유사도 추출을 최종적으로 수행한 결과 값을 저장하는 테이블로서, 사용자별로 키워드 간 연관성이 1회 이상 있는 모든 사용자간의 키워드 유사도를 Map/Reduce 프로그래밍을 통한 분산처리를 이용하여 도출 한 뒤 저장한다.

표 8. 블로그 사용자간 키워드 유사도 결과 저장 테이블 [USER_SIMILARITY])

Row Key		블로그 UID
Column Family:USERS	[USER1]	UID와 [UID1]간의 키워드 유사도값
	…	…
	[USERn]	UID와 [UIDn]간의 키워드 유사도값

2.4 Map/Reduce 설계

본 논문은 대용량의 블로그 연결망 정보를 분석하기 위해 구글이 2004년도에 소개한 Map/Reduce 기술을 이용하도록 한다[8]. MapReduce는 Map과 Reduce라는 2가지 방법을 조합해서 데이터를 처리하는 기술이다. Map은 어떤 데이터의 집합을 받아들여 새로운 데이터를 생성하는 프로세스이며, Reduce는 Map에 의해 만들어진 데이터를 모아서 최종적으로 원하는 결과로 만들어 내는 프로세스다.

2.4.1 블로그 간 연결정보 분산처리

블로그의 연결망을 분석하기 위해 1차적으로 명시적인 블로그 간 연결 정보를 분석할 수 있어야 한다. [USER_TRANSACTION] 테이블에 저장되어 있는 블로그 연결망 정보 데이터를 분산처리하기 위한 Map/Reduce 설계는 아래와 같다.

표 9. 블로그 간 연결정보 분석 Map/Reduce 설계

Map	Input Data<K,V>	<Seq,Row>
	Output Data<K,V>	<UID,{ActionType,Verb,UID,ROW}>
Reduce	Input Data<K,I<V>>	<UID,Iterator<{ActionType,Verb,UID,ROW}>
	Output Data<K,V>	<UID-UID,Score>

Map의 Input으로는 [USER_TRANSACTION] 테이블의 일련번호(Row Key)가 Key로, 실제 ROW 정보가 Value로 들어오도록 설계하며, 입력된 데이터를 주체, 피 주체로 이분화 한 뒤 다시 Output의 Key로 분석 주체 블로그 UID를 Value로 주체, 피 주체 구분 플래그, 행위동사, 분석 대상 블로그 UID, ROW 정보를 넘기도록 한다.

이렇게 Map의 작업으로 모인 데이터를 Reduce 단계에서는 Iteration을 수행하면서 2.1절의 가중치 설계 테이블에 설정된 값에 따라 모든 연결정보 점수를 합산하게 되며, 시간의 경과에 따른 가중치 보정을 위해 로지스틱 성장곡선을 이용한다. 이 과정을 통해 블로그 간 명시적인 연결행위에 대한 연결망 점수를 추출할 수 있게 된다.

2.4.2 블로그 별 키워드, 키워드별 블로그 통계 분산처리

Map의 입력데이터로 블로그 별 게시물정보를 파싱한 뒤 키워드를 추출하여, 첫 번째로 블로그 UID 와

키워드를 Key로, Value로는 정수 1을 Output 하도록 하고, 두 번째로 키워드를 Key로 블로그 사용자 UID를 Value로 Output 하도록 하며, 구분을 위해 전자는 “001”, 후자는 “002” 이라는 플래그 변수를 키 값 앞쪽에 두도록 한다. Map의 작업결과로 모인 데이터를 Reduce 단계에서는 Key 값의 앞 3 바이트를 파싱하여 “001”인 경우 단순히 Iteration 을 수행하면서 합산한 뒤 블로그 사용자별 키워드 통계정보를 Output 하도록 하고, “002”인 경우에는 키워드별 블로그 사용자 정보를 바로 [KEYWORD_USERS] 테이블에 기록하도록 한다.

표 10. 블로그 별 키워드, 키워드별 블로그 통계 추출 Map/Reduce 설계

Map	Input Data<K,V>	<Seq,Row>
	Output Data<K,V>	<001-UID-Keyword,Integer(1)> 혹은, <002-KEYWORD,UID>
Reduce	Input Data<K,I<V>>	<001-UID-Keyword,Iterator <Integer(1)>> 혹은 <002-KEYWORD,Iterator<UID>>
	Output Data<K,V>	<UID-Keyword,TF>

2.4.3 블로그 사용자간 키워드 유사도 분산 처리

2.4.2절의 Map/Reduce 분산처리 결과로 블로그 사용자별 Term Frequency, 키워드별 사용자 집합을 구할 수 있다. 다음 단계는 수집된 자료를 기반으로 실제 블로그 사용자별 키워드 유사도 계산을 위한 분산 처리 모델을 설계하도록 한다.

Map의 입력데이터로는 [KEYWORD_USERS] 테이블의 ROW 값을 받아들이게 한다. 이때 Key는 키워드, Value로는 블로그 사용자 UID가 LIST 형식으로 들어오게 되는데, 이때 모든 사용자간 연결 가능한 경우의 수를 Output 하여야 한다.

Reduce 단계에서는 키워드유사도 분석대상 블로그 사용자가 Key로 입력되게 되며, Value로는 서로 간에 1번 이상 중복되는 키워드가 모여지게 된다. 이 단계에서 [USER_KEYWORD] 테이블에 저장된 사용자별 Term Frequency 정보를 Vector화 한 뒤 코사인 유사도 공식을 이용하여 블로그 사용자별 키워드 유사도 정도를 구함으로써, 블로그 속성을 대표할 수 있는 키워드를 통한 블로그 간 연결망 정도의 점수를 추출 할 수 있다.

표 11. 블로그 간 키워드 유사도 분석 Map/Reduce 설계

Map	Input Data<K,V>	<Keyword,LIST<UID>>
	Output Data<K,V>	<UID-UID,Keyword>
Reduce	Input Data<K,I<V>>	<UID-UID,Iterator<Keyword>>
	Output Data<K,V>	<UID-UID,Double<Similarity>>

Ⅲ. 연결망 분석 시스템 설계 및 실험

3.1 연결망 분석 시스템 구성

그림 1은 블로그 연결망 분석 시스템의 구조도 이다. 시스템은 다음과 같이 6개의 모듈로 구성된다.

(1) 웹 정보 수집기(Web Crawler): 인터넷에 산재되어 있는 웹문서를 수집하는 역할을 담당한다. 운용자가 지정한 첫 엔트리 포인트 웹페이지에서부터 수집을 시작하여, 해당 페이지에 포함된 모든 링크를 순회하면서 다른 웹 웹문서를 수집한다. 수집방식은 깊이우선방식으로 최대 깊이가 100 이상일 경우에는 무시하도록 구성한다.

(2) 사용자 정의 블로그 파서(Custom Blog Parser): 수집된 웹 문서 중에서 블로그 문서만을 추출하는 기능을 담당한다. 또한, 블로그 문서로 판단되는 경우 해당 블로그의 Home URI, 게시물 제목, 게시물 내용, 키워드(태그)와 같은 문서의 메타정보를 추출하고, 블로그 간 서로 명시적으로 연결되어 있는 행위를 파싱하여 연결망 기초 분석 데이터로 등록한다.

(3) 환경 설정 모듈(Configuration Module): 연결망 분석에 필요한 가중치 정보들을 저장 관리 하는 기능을 담당하고 있다. 웹 인터페이스를 통해 운용자가 입

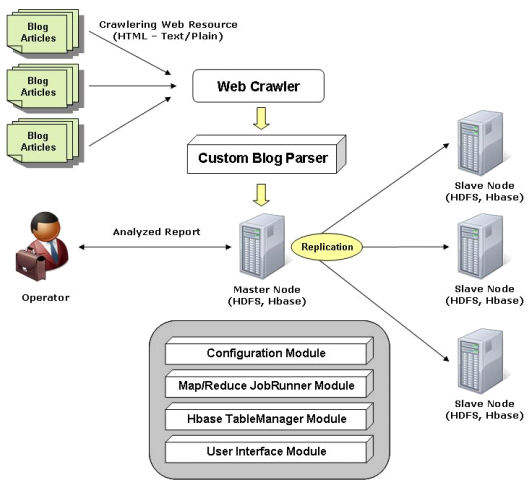

그림 1. 블로그 연결망 분석 시스템 구조도

력한 시스템 환경설정에 대한 값을 XML 파일로 로컬에 저장한 뒤, HDFS(Hadoop Distributed File System)에 업로드 하여 분산노드(Slave Node)에서도 해당 환경설정 값을 읽어 들여서 분석처리에 이용되도록 한다.

(4) Map/Reduce 분산처리 작업 구동 모듈(Map/Reduce Job Runner Module): 2.4절에 설계한 각 분산처리 행위의 Mapper, Reducer Class를 구동시키는 역할을 담당하고 있다. 이 모듈은 또한 환경 설정 모듈에 설정된 Task 수에 따라, 유동적으로 복수의 Slave Node로 분산처리를 다시 나누어서 작업을 수행하도록 한다.

(5) HBase 테이블 관리 모듈(HBase Table Manager Module): 데이터 입력, 수정, 삭제를 제어한다. Map 작업 수행 시 환경 설정 모듈에 설정된 Task 수에 따라 적절하게 데이터를 분리하여 Slave Node에 할당하도록 하며, Reduce 과정을 통해 연결망 분석이 끝난 데이터를 저장하도록 한다.

(6) 사용자 인터페이스 모듈(User Interface Module): 사용자와의 인터페이스를 구성해주는 웹 기반의 모듈이다. 블로그 연결망 시스템의 환경설정에 필요한 일련의 변수 값을 정의 할 수 있으며, 분석이 끝난 연결망 데이터를 확인할 수 있다.

3.2 실험 및 결과

Map/Reduce 분산처리 시스템은 하드웨어 사양이 모두 동일한 1대의 Master Node와 3대의 Slave Node로 구성하였으며, 분석시스템 간 네트워크 대역폭은 1Gbps로 설정하여 실험하였다. 실험은 이글루스(Egloos), 티스토리(Tistory) 블로그 포털 하에 운영되는 약 1000개의 개인 블로그를 수집대상으로 하였으며, 약 10일 동안 수집된 총 25,000건의 게시물과 114,000건의 블로그 간 연결정보를 연결망 분석 시스템의 분석기초 데이터로 설정하였다.

위의 환경에서 실험한 결과, 기존의 Neighbor 분석 방식이나 Centrality 분석 방식인 경우 452개(45%)의 노드가 구성되었으나, 본 논문에서 제시한 키워드 유사도가 반영된 연결망 분석 방법에서는 954개(95%)의 노드가 분석노드로 검출되었다. 이는 제안 방식이 명시적으로 연결된 노드간의 관계가 없더라도, 블로그 속성을 대표할 수 있는 키워드를 기반으로 묵시적인 연결망을 구성함으로서 기존의 분석방법론보다 더 많은 노드 간의 관계를 추출할 수 있음을 의미한다.

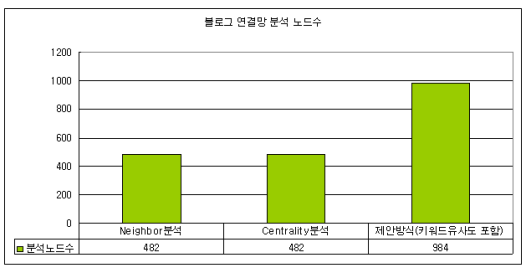

그림 2. 분석방식별 연결망에 포함된 노드 수

아래 그림 3은 특정노드와 연결되어 있는 노드 A, 노드 B의 시간의 흐름에 따른 연결강도를 선 그래프로 나타낸 것이다.

Neighbor 분석방식의 경우 노드 A, 노드 B의 연결강도가 시간이 60일까지 흘렀음에도 불구하고 계속 32로 유지되는 반면, 본 논문에서 제시한 분석방식의 경우 노드 A의 연결강도는 처음에 28, 60일이 지났을 때에는 6의 연결강도를 가지고, 노드 B의 연결강도는 처음에 28, 60일이 지났을 때에는 19로 시간에 따라 떨어짐을 알 수 있다.

이는 Neighbor 분석방식의 경우 시간의 경과에 따른 연결 강도의 변화고려가 없는 반면, 본 논문에서 제시한 방법론의 경우 각 노드의 연결 행위에 따라 시간의 변화에 따른 가중치 변동으로 인해 최근의 행위가 과거의 행위보다 우선시되는 연결망 분석이 가능함을 의미한다.

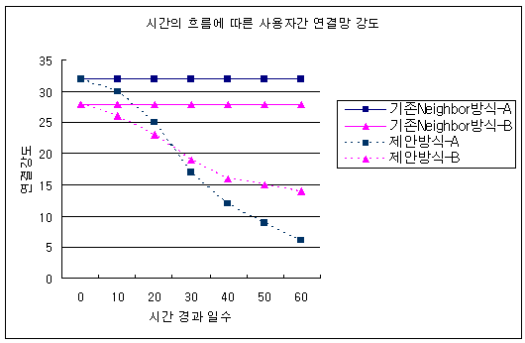

그림 3. 분석방식별 시간의 흐름에 따른 연결망 강도 변화 그래프

Ⅳ. 결 론

본 논문에서는 Map/Reduce를 이용하여 블로그 연결망을 분석하는 시스템의 설계에 대하여 논의 하였다. 제안하는 연결망 분석 시스템은 기존의 Neighbor 분석 방식과 달리, 행위시간의 지남에 따른 보정 수식을 적용하여, 연결망 성향정도의 변화추이 분석이 가능하도록 하였으며, 또한 Centrality 분석 방식과는 다르게 실

질적인 연결행위가 없더라도, 컨텐츠의 유사도를 이용하여 각 블로그 사용자간 묵시적인 연결망을 포함한 분석이 가능하였다. 마지막으로 대용량의 연결망 기초자료 데이터를 Big-Table에 저장하고 Google의 Map/Reduce 프로그래밍 모델을 이용하여 효율적으로 분산처리 분석이 가능함을 입증하였다.

참 고 문 헌

〔1〕 S. Wasserman and K. Faust, “Social Network Analysis: Methods and Applications”, Cambridge University Press, 1994

〔2〕 L. Adamic, O. Buyukkokten, and E. Adar, ″A Social Network Caught in the Web″ Frist Monday, Vol.8, No.6, pp.1-22, 2003

〔3〕 X. Song et al., ″Mining in Social Networks Information Flow Modeling based on Diffusion Rate for Prediction and Ranking″, Proc. Int’l. Conf. on World Wide Web, pp.191-200, 2007

〔4〕 J. Iribarren and E. Moro, ″Information Diffusion Epidemics in Social Networks″, Arxiv, 2007

〔5〕 ㈜ 다음 커뮤니케이션, http://www.tistory.com

〔6〕 ㈜ SK Communications, http://www.egloos.com

〔7〕 A. Chin and M. Chignell, “A Social Hypertext Model for Finding Community in Blogs”, Proc. Int’l. Conf. on Hypertext and Hypermedia, pp. 11-22, 2006

〔8〕 J. Dean and S. Ghemawat, ″MapReduce: Simplified Data Processing on Large Clusters″, 6th Symposium on Operating System Design and Implementation, Google Inc., 2004

〔9〕 김경희, 배진아, “30대 블로거들의 블로그 매개 커뮤니케이션 연구”, 한국언론학보, 제50권 제5호, 2006

조 인 휘 (Inwhee Joe) 정회원

1983년 2월 한양대학교 전자공학과

1994년 12월 미국 University of Arizona, Electrical and Computer Engineering, M.S.

1998년 9월 미국 Georgia Tech, Electrical and Computer Engineering, Ph.D.

1992년 12월 (주) 데이콤 종합연구소 선임연구원

2000년 6월 미국 Oak Ridge 국립연구소 연구원

2002년 8월 미국 Bellcore Lab (Telcordia) 연구원

2002년 9월~현재한양대학교 컴퓨터공학부 부교수

<관심분야> Mobile Internet, Cellular System and PCS, Sensor Networks, Mobility Management