http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.20, No.2, May 2015, pp.155-166
http://dx.doi.org/10.7838/jsebs.2015.20.2.155

부호 그래프에서의 빠른 랜덤워크 기법

Fast Random Walk with Restart over a Signed Graph

명재석(Jaeseok Myung)*, 심준호(Junho Shim)**, 서보밀(Bomil Suh)***

초 록

랜덤워크는 그래프 기반의 랭킹 기법들에서 빈번히 사용되지만, 그래프 간선에 음수 가중치를 가지는 부호 그래프는 고려하지 않는다. 이 논문에서는 하이더의 균형 이론을 적용하여 랜덤워크 수행 시 음수 가중치를 처리하는 기법을 제안한다. 제안 기법은 추천 시스템에 적용되었으며, 사용자가 선호하지 않는 아이템을 걸러내는 데 효과가 있음을 실험을 통해 보인다. 제안한 모델의 성능을 위해 기존의 Top-k 랜덤워크 계산 기법인 BCA를 확장한 Bicolor-BCA 알고리즘을 제안한다. 제안 알고리즘은 임계값이 필요한데, 실험을 통해 임계값에 따른 정확도와 성능의 변화를 살펴본다.

ABSTRACT

RWR (Random Walk with Restart) is frequently used by many graph-based ranking algorithms, but it does not consider a signed graph where edges may have negative weight values. In this paper, we apply the Balance Theory by F. Heider to RWR over a signed graph and propose a novel RWR, Balanced Random Walk (BRW). We apply the proposed technique into the domain of recommendation system, and show by experiments its effectiveness to filter out the items that users may dislike. In order to provide the reasonable performance of BRW in the domain, we modify the existing Top-k algorithm, BCA, and propose a new algorithm, Bicolor-BCA. The proposed algorithm yet requires employing a threshold. In the experiment, we show how threshold values affect both precision and performance of the algorithm.

키워드 : 랜덤워크, 균형이론, 추천, 부호 그래프
Random Walk, Balance Theory, Recommendation, Signed Graph

이 논문은 2015 한국전자거래학회 춘계학술대회 우수논문으로 선정되어 수정 확장된 것임. 본 연구는 숙명여자대학교 교내연구비지원에 의해 수행되었음(과제번호 1-1406-0005).

* First Author, Samsung Electronics Co., Korea(jsmyung@europa.snu.ac.kr)
** Corresponding Author, Division of Computer Science, Sookmyung Women's University, Korea (jshim@sookmyung.ac.kr)
*** Co-Author, Division of Business Administration, Sookmyung Women's University, Korea (bmsuh@sookmyung.ac.kr)

Received: 2015-05-08, Review completed: 2015-05-19, Accepted: 2015-05-20

1. Introduction

Thanks to its flexibility to accept various types of data, the graph data model is used in many applications [11]. Especially, the graph-based recommendation system is a well-known application that can recommend users some items by measuring the proximity between nodes. For instance, if we regard users and items as nodes and then convert the purchase history into edges between nodes, it can sort out other items which are linked through the purchase history of users based on proximity.

Random Walk with Restart (RWR) is one of the most representative proximity measuring techniques which are typically used in the graph-based application system. A representative example of RWR application in the recommendation and search field is Google's PageRank [13]. The PageRank algorithm models documents as nodes and defines hyperlinks between them as edges, and calculates the probability of a random surfer staying on a certain page at a certain time. Since the success of this approach, other expanded models which overcome the shortcomings of the early model have been proposed. Personalized PageRank [7], SimRank [10], and ObjectRank [3] are the typical standard graph-based proximity measures.

In this paper, we consider a graph that the RWR model cannot process, and discuss a technique that can handle this type of graph. The type of graph we handle in this paper is a signed graph. The weighted values of edges of the existing random walk model are real numbers between 0 and 1, and if the weights of all edges that belong to one node are added up, the sum amounts to 1. This shows that the RWR model is based on the calculation of probability. Therefore, if the edge weight has a negative value, the basic hypothesis of RWR does not hold true, and therefore it cannot process a signed graph.

A singed graph can appear in many application services. For instance, let us consider a movie rating service with the scale of 1 to 5 points. If we look into nodes between movies and users, the weighted values of 1 to 5 points are given to edges between them. Generally, the point between 1 and 2 means a negative review of a user about a movie. Therefore, the proposed graph will be modified to have the weighted values of -2 to 2 points. Likewise, singed graphs often occur in the kind of services which deal with the feedbacks of users, and this can be utilized as important information for the recommended system.

The basic method to process signed graphs is the shifting technique which convert the graphs of -2 to 2 points to those of 1 to 5 points. However, this technique places priority on items that users dislike (those with points of 1 or 2) over those which they have not used before (those not linked with edges). A simple shifting technique is not enough to

solve problems, because it is important for a recommendation system to reduce the number of items that users dislike.

In this study, we propose a novel method to perform Random Walk with Restart over signed graphs. The major content is as follows.

- Based on Balance Theory by F. Heider, which is studied in social science, we propose a technique to process signed graphs. This model is called Balanced Random Walk.
- For a fast calculation of our proposed model, we propose a new algorithm by expanding the exiting Top-K RWR algorithm. This algorithm is called Bicolor-BCA Algorithm.
- We perform the experimental evaluation to show the effectiveness of the proposed algorithm, and also how the threshold values can affect the algorithm.

The rest of this paper is as follows. In Section 2, we introduce the background knowledge and related studies and offer the problem definition. In Section 3, we introduce Balanced Random Walk. Section 4, we introduce Bicolor-BCA Algorithm for a faster calculation of Balanced Random Walk. In Section 5, we verify our proposed algorithm through experiments. Finally in Section 6, we present the conclusion and future study direction.

2. Backgrounds

2.1 Random Walk with Restart Model

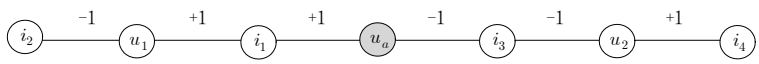
In Random Walk with Restart (RWR), the steady-state probability ($\overrightarrow{\pi_t}$) of a random surfer of each node on a given graph $G=(V, E)$ is defined as the following equation, and the computation of $\overrightarrow{\pi_t}$ must be repeated until the condition ($\overrightarrow{\pi_t} \cong \overrightarrow{\pi_{t-1}}$) is met [15].

$$\overrightarrow{\pi_t} = cT\overrightarrow{\pi_{t-1}} + (1-c)\vec{q}$$

In the above equation, T means the transition matrix of a graph, $\vec{q}$ is a vector where the value of a query node is 1 and the rest are 0. c is a constant number of 0.85, which is used as a factor to determine whether it will transmit to adjacent nodes or return to the query node.

2.2 Signed Graphs Processing Technique

The typically used methods to process signed graphs are the shifting method and the splitting method. To explain the methods, let us consider a simple graph as shown in



〈Figure 1〉 Example of a Signed Graph

<Figure 1>.

We will measure the proximity of items ($i_1 \sim i_4$) based on u_a which is located at the center.

The first method is the shifting method. In this method, we substitute the values of -1 to 1 with the corresponding values of 1~3. The results are as follows in <Figure 2>. Based on the condition which gives higher scores to nodes at the same distance from the adjacent nodes, the results of the arrangement will be in the order of $i_1 > i_3 > i_2 > i_4$. In accordance with the definition of RWR, it is general to give a higher score to more adjacent nodes. As a result, higher score is given to i_3, which has a negative score in the original graph.

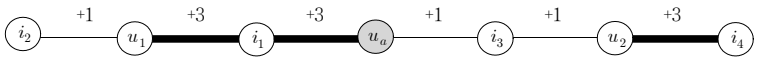
The second available method is the splitting method [6]. The splitting method divides one graph $G=(V, E)$ into two sub-graphs: $G^+=(V, E^+)$ and $G^-=(V, E^-)$ (<Figure 3>). If we perform the RWR respectively on each split sub-graph, we can get their respective probability distribution of $\vec{\pi}^+$ and $\vec{\pi}^-$. Finally, we can use the value of $(\vec{\pi}^+ - \vec{\pi}^-)$ as the ranking function.

Although the splitting method can reflect both the negative and positive meaning of edges, the connection relationship between nodes often disappears in the middle of the process. It has a shortcoming, in that it cannot generate any recommendation results in some cases. Therefore, a new method is required to make up for such disadvantage.

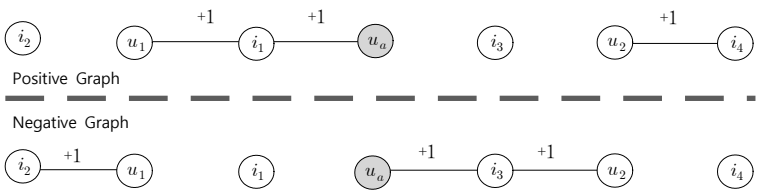
2.3 Balance Theory

Balance Theory is a social science theory introduced by Fritz Heider, which argues that if an imbalance occurs in a cognitive aspect, individuals try to continue to maintain the balance by changing their attitudes [9]. The balance state can be explained intuitively with three nodes over graphs.

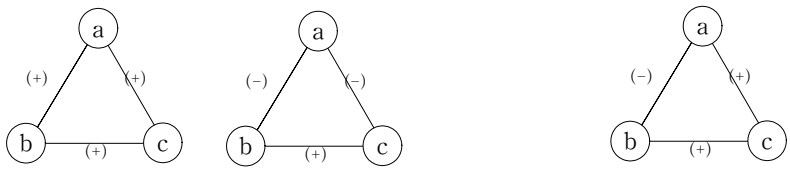
Let us consider that the graph in <Figure 4> shows the relationship of a friend and an enemy. '+' marked on edges between nodes



〈Figure 2〉 Applying Shifting Method to a Signed Graph



〈Figure 3〉 Applying Splitting Method to a Signed Graph



〈Figure 4〉 Balanced or Unbalanced Graphs by Balance Theory

means a friendly relation between nodes, while ' - ' signifies a hostile relation between nodes. Nodes 'a,' 'b' and 'c' in the leftmost figure all have friendly relationships with each other and are in a balanced state. The relationship between nodes 'b' and 'c' in the second leftmost figure is friendly, while their relationships with 'a' are hostile, which can be also said to be in a balanced state. On the rightmost figure, node 'c' has a friendly relationship with nodes 'a' and 'b,' while the relationship between 'a' and 'b' is hostile. In such a case, an imbalanced state occurs. According to Balance Theory, if an imbalance happens, it tries to return to a balanced state by changing attitudes. In a balanced state, the following four propositions hold true. (1) The friend of my friend is my friend. (2) The enemy of my friend is my enemy. (3) The friend of my enemy is my enemy. (4) The enemy of my enemy is my friend.

We propose Balanced Random Walk applied the above Balance Theory. Of course, it is impossible to apply Balance Theory to all kinds of application services. But the theory can play a critical role in such domains as product recommendation service and social network analysis.

2.4. Top-k RWR Algorithms

If we use the RWR model, we can calculate the proximity of a certain query node to all nodes. However, if the graph-based recommendation system can obtain the number ('*k*') of items which are to be recommended to real users, the rest of the calculation is meaningless. Studies were actively conducted on the calculation (the Top-k RWR).

Algorithms in the first category exploit Monte-Carlo technique [1, 2]. As the RWR model calculates the probability that a random user will stay at a node at a certain time, Monte-Carlo technique is the intuitive materialization of the model which allows a multiple number of random surfers to simultaneously perform the RWR by intuitively implementing the RWR model. However, if the number of random surfers surpasses that of edges, this model requires unnecessary calculation. To overcome this shortcoming, BCA (Bookmark Coloring Approach) was introduced, which can reduce unnecessary calculation by transmitting the probability values [5]. And the last category

is the Equation Solving Approach to simplify the entire calculation by using the inverse of matrices [8]. However, this method has its own disadvantage, in that if a graph changes, it is necessary to recalculate the inverse of matrices again.

In this study, we developed Balanced Random Walk by expanding the BCA algorithm among the above three methods. The detailed procedures will be explained in Section 4.

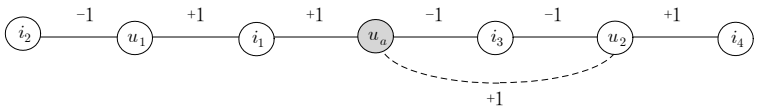
3. Balanced Random Walk

Let us reconsider the example mentioned in Section 2 to implement Balanced Random Walk (BRW). In accordance with Balance Theory by F. Heider, we presume that a graph is in a balanced sate, and then we perform the RWR to obtain the results in the order of $i_1 > i_4 > i_2 > i_3$. To this end, we conceptually create a virtual edge between 'u_a' and 'u_2', both of whom dislike the same item 'i_3', and form a friendly relationship between them. After calculating the respective scores of the positive and negative relationship separately, we reflect the difference in the score of the concerned node (<Figure 5>).

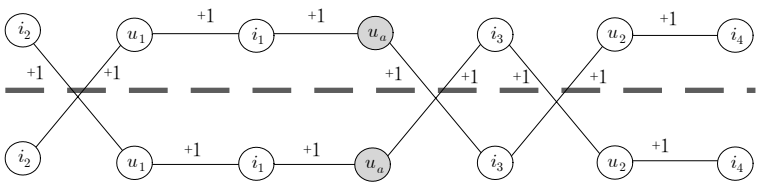
We obtain a new graph (G') by undergoing the following procedures on the given graph ($G = (V, E)$) in a more ordinary form.

- Copy the graph (G) and obtain G^+ and G^-($G = G^+ = G^-$). The union ($G^+ \cup G^-$) of these can be defined as the early (G').
- In G^+ and G^-, the edge which has a negative weight is changed into the cross edge between them.

As the RWR is performed on G', the query node in an early stage starts with u_a of G^+. In the above example, although G^+ and G^- are not connected and separated from each other, the correlation between the sub-graphs usually exists in case of a big graph. For example, if there exists 'u_3', who likes both 'i_3' and 'i_4', the two sub-graphs will be con-



〈Figure 5〉 Virtual Edge in a Signed Graph



〈Figure 6〉 Applying BRW to a Signed Graph

nected. Therefore, vectors $\vec{\pi}^{+}$ and $\vec{\pi}^{-}$ are created as the results of the RWR. These values can be utilized to identify the preference of 'u_a.' Finally we use $(\vec{\pi}^{+} - \vec{\pi}^{-})$ as the value of the ranking function.

In conclusion, if Balance Theory is applied to a graph, it is possible to measure the proximity based on our proposed BRW. The graphs in which Balance Theory holds true are exceptionally in a great number. The friend relationship graph of a social network is a typical graph which Balance Theory is applied to. Most of the cases are that the friend of my friend is my friend. However, if the enemy of my enemy is my enemy, the balance theory does not always hold true, to which much attention should be paid. If we take another example, in case that a graph shows a correlation between nodes, Balance Theory holds true. If negative relationships overlap, it is transformed into a positive correlation. We conduct an experiment on a movie recommendation algorithm by using movie review grades. In this case, our proposed method has a positive influence on the recommendation results, which will be explained in detail in Section 5.

4. Bicolor-BCA Algorithm

As mentioned in Section 1, the random walk model performs a number of calculations to obtain a converged probability. Especially because it repetitively performs the computation of matrix multiplication, which requires a large amount of calculations, it is not suitable for search or recommendation applications. Because of that, there is a service which does pre-computation and offers the results when it has an inquiry. This kind of service has its own shortcoming, in that it cannot well respond to a change in a graph.

BCA (Bookmark Coloring Approach) is an algorithm which returns the corresponding Top-k result node to the given query node [4]. This algorithm is implemented by mimicking the phenomenon of smearing of paint to neighboring area on paper. That is to say, if the amount of '*1*' of coloring agent is injected into a query node, the corresponding coloring agent (*1-c*) remains in the current node, and the rest of paint ('*c*') will smear into the neighboring nodes. The process will be performed in a recursive manner, and the algorithm comes to stop when the amount of the remaining coloring agent to smear into neighboring nodes becomes under the threshold value, *theta* (θ). Finally, the remaining coloring agent of each node is used as the score of the concerned node.

As we defined a model over a new type of graph, we could not use the existing methods. <Figure 7> shows the pseudo code of Bicolor-BCA algorithm which performs the Balanced Random Walk on a signed graph.

Input: A graph $G=(V,\ E=E^{+}\cup E^{-})$, a query node $n\in V$
Output: A score vector $\vec{\pi}$
1. Insert $q=(n,\ 1)$ to Q^{+}
2. At each step, we compare top elements in Q^{+} and Q^{-}, and select the bigger element $q=(n,\ q)$
   • If $p<\theta$, then stop to deque.
   • If $q\in Q^{+}$, then $\vec{\pi}_{n}^{+}+=(1-c)\times p$
   • Otherwise, $\vec{\pi}_{n}^{-}+=(1-c)\times p$
   • Propagate $c\times p$ to neighbors
     - If $q\in Q^{+}$ and $e\in E^{+}$, then insert new element to Q^{+}
     - If $q\in Q^{-}$ and $e\in E^{+}$, then insert new element to Q^{-}
     - If $q\in Q^{+}$ and $e\in E^{-}$, then insert new element to Q^{-}
     - If $q\in Q^{-}$ and $e\in E^{-}$, then insert new element to Q^{+}
3. Ranking score = $(\vec{\pi}^{+}-\vec{\pi}^{-})$

〈Figure 7〉 Bicolor-BCA Algorithm in Pseudo Code

The algorithm uses two priority queues (Q^{+} and Q^{-}), and the score of nodes are stored in $\vec{\pi}^{+}$ and $\vec{\pi}^{-}$. The query node is 'n'. At an early stage of algorithm, we added a value to Q^{+} to enable random suffers to move. And then, after extracting the value of the node with the largest remaining amount of paint among queues, we multiply the node value by $(1-c)$. As a result, the amount of remaining transmissible paint remains as much as $c\times p$, and the value is transmitted to the next appropriate queue depending on the current queue state and on the signs of edges. When the amount of remaining coloring agent falls below a certain point, the algorithm comes to end. Finally, $(\vec{\pi}^{+}-\vec{\pi}^{-})$ is used as the score of the concerned node.

One of the advantages of Bicolor-BCA algorithm is a low space complexity. In Section 3, BRW has its own shortcoming, in that the number of nodes and edges conceptually need to be doubled in order to implement the algorithm. Considering that a massive graph often occurs in the recent years, it is too bulky. However, the proposed algorithm can be implemented with two priority queues and two score vectors. Each of them can have a length equivalent to the number of nodes or a maximum number of 'n'. Therefore, it can be performed at a relatively low space complexity.

5. Experiment

Through the experiment, we verified the ranking accuracy of our proposed BRW and also the computational efficiency of Bicolor-BCA algorithm. The data used in the experiment comes from MovieLens-100K [12]. This is a dataset that contains 1,682 movies reviews of 943 users and 100k number of rating data, and is often used to verify the performance of a recommender system.

The collaborative filtering (CF) method is used as a comparative model against the proposed model. We used the user-oriented CF [14]. As other comparative object model, we carried out the shifting method and the splitting method using the Random Walk model

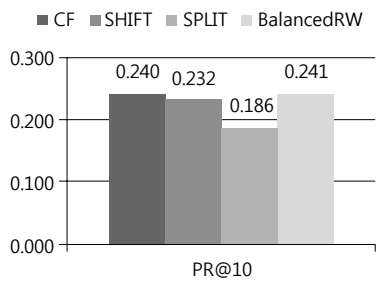
[6]. The conversion method was used to change the values of 1 to 5 points to those of -2 to 2. In the splitting method, the scores of 1 to 2 mean negative opinions, while the scores of 4 to 5 mean positive opinions. As the original BCA algorithm cannot process negative edge, it is not suitable for objective comparison. Thereofore, it was excluded from the experiment.

To check the recommendation accuracy, Precision (PR@10) and Inverse Precision (IPR@10) are used. PR@10 shows how many items which receive positive reviews (average rating score of 4 and higher) are distributed among 10 recommendation results. Likewise, IPR@10 measures how many items which receive negative reviews (average rating score of 2 and lower) are distributed among 10 recommendation results. Therefore, if the IRP is high, the reliability of the recommendation results is very low.

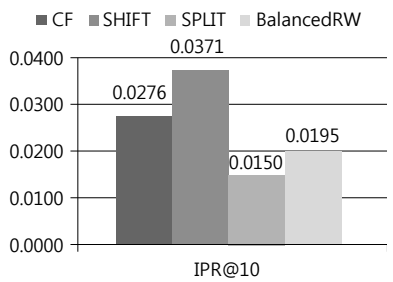
<Figure 8> shows the results of the experiment to verify the effect on accuracy. In terms of PR@10, our proposed algorithm shows a higher level of accuracy than Shift and Split methods. This means that our proposed system overcomes for the shortcoming of the Shift method by correcting the wrong interpretation of negative opinions, while it also make up for the disadvantage of the Split method that it sometimes fails to generate the recommendation results. Also, the accuracy of our proposed system does not fall behind compared to CF, which is commonly used. Generally as a graph-based recommendation system is not subject to limitation depending on the type of recommendation items, it has the merit to be used to implement a more flexible recommendation system.

In terms of IPR, we can see more distinctive differences compared to CF. The proposed algorithm has a lower probability of including those items that users dislike, compared to other algorithms. Although the Split method showed relatively lower results than the proposed algorithm, it sometimes fails to generate any recommendation results.

In addition, we perform the experiment on the efficiency of Bicolor-BCA method (<Figure 9>). By changing the threshold among a dataset



(a) accuracy w.r.t. affirmative recommendation: PR(10)



(b) accuracy w.r.t. negative recommendation: IPR(10)

〈Figure 8〉 The Performance of Bicolor-BCA Compared to Others

of the same MovieLens, we measure changes in accuracy. If the threshold is low, a lower amount of coloring agent can smear into a wider neighboring area, which results in more frequent dequeue computations and requires more time. If we look at the following results, the time required to process an inquiry at $\theta = 10^{-7}$, it takes almost three seconds. Given this, it is too slow to handle a large dataset in a real application. At $\theta = 10^{-6}$, it shows an accuracy level similar to the CF level in the aforementioned experiment.

θ	Time (ms)	Dequeue	PR@10
10^{-5}	38	12,730	0.159
10^{-6}	247	130,817	0.241
10^{-7}	2,820	1,270,424	0.264

〈Figure 9〉 The Performance of Bicolor-BCA by Different Threshold Values

6. Conclusions

We proposed BRW (Balanced Random Walk) and Bicolor-BCA algorithm to process signed graphs, in order to overcome the shortcomings of the existing RWR model. The BRW model is designed to allow random surfers to move in accordance with Balance Theory by F. Heider, and it shows a higher accuracy than the user-oriented collaborative filtering of a real recommendation system.

The values of BRW model can be quickly computed over the Top-k items by Bicolor-BCA algorithm. By changing the threshold value, we can select higher accuracy or faster performance time. In the future study, we need to produce a significant result by applying our proposed algorithm to the analysis of a correlation coefficient graph where Balance Theory holds true. Another important future study is to develop another model that can process signed graphs, except for Balance Theory, and to perform a comparable study with the model presented in this paper.

References

[1] Avrachenkov, K., Litvak, N., Nemirovsky, D. A., Smirnova, E., and Sokol, M., "Monte Carlo Methods for Top-k Personalized PageRank Lists and Name Disambiguation," INRIA Research Report No.7367, 2010.

[2] Bahmani, B., Chakrabarti, K., and Xin, D., "Fast Personalized PageRank on MapReduce," Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, 2011.

[3] Balmin, A. and Hristidis, V., "ObjectRank: authority-based keyword search in databases," Proceedings of the Thirtieth international conference on Very large data bases, 2004.

[4] Berkhin, P., "Bookmark-coloring approach to personalized pagerank com-

puting," Internet Mathematics, 2006.

[5] Chakrabarti, S., Pathak, A., and Gupta, M., "Index Design and Query Processing for Graph Conductance Search," The VLDB Journal, 2011.

[6] Clements, M., Vries, A. P., and Reinders, M. J. T., "Exploiting Positive and Negative Graded Relevance Assessments for Content Recommendation," Algorithms and Models for the Web-Graph, Lecture Notes in Computer Science, 2009.

[7] Fogaras, D., Rácz, B., Csalogány, K., and Sarlós, T., "Towards Scaling Fully Personalized PageRank: Algorithms, Lower Bounds, and Experiments," Internet Mathematics, 2005.

[8] Fusiwara, Y., Nakatsuji, M., Yamamuro, T., Shiokawa, H., and Onizuka, M., "Efficient Personalized PageRank with Accuracy Assurance," Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data minin, 2012.

[9] Heider, F., The Psychology of Interpersonal Relations, John Wiley & Sons, 2013.

[10] Jeh, G. and Widom, J., "SimRank: a measure of structural-context similarity," Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002.

[11] Kang, S. and Shim, J., "Empirical Analysis on the Shortcut Benefit Function and its Factors for Triple Database," Journal of Society for e-Business Studies, Vol 19, No 1, Society for e-Business Studies, 2014.

[12] MovieLens, http://grouplens.org/datasets/movielens/.

[13] Page, L., Brin, S., Motwani, R., and Winograd, T., "The PageRank Citation Ranking: Bringing Order to the Web," Technical Report, Stanford InfoLab, http://ilpubs.stanford.edu:8090/422/.

[14] Su, X. and Khoshgoftaar, T. M., "A Survey of Collaborative Filtering Techniques," Journal Advances in Artificial Intelligence, 2009.

[15] Tong, H., Faloutsos, C., and Pan, J., "Fast Random Walk with Restart and Its Applications," http://repository.cmu.edu/compsci/537/, 2006.

저 자 소 개

명재석 (E-mail: jsmyung@europa.snu.ac.kr)
2007년 성균관대학교 정보통신공학부 졸업 (학사)
2014년 서울대학교 컴퓨터공학부 졸업 (석박사 통합과정)
2014년~현재 삼성전자 디자인경영센터 재직 (책임)
관심분야 데이터베이스, 전자상거래, 추천시스템

심준호 (E-mail: jshim@sookmyung.ac.kr)
1990년 서울대학교 계산통계학과 졸업 (학사)
1994년 서울대학교 계산통계학과 전산과학전공 (석사)
1998년 Northwestern University, Electrical & Computer Engineering (박사)
2001년~현재 숙명여자대학교 컴퓨터과학부 교수
관심분야 데이타베이스, 전자상거래, 상품정보, 온톨로지

서보밀 (E-mail: bmsuh@sookmyung.ac.kr)
1994년 KAIST 전산학과 (학사)
1997년 KAIST 경영공학 (석사)
2003년 KAIST 경영공학 (박사)
2002년~2004년 LG CNS 선임컨설턴트
2004년~현재 숙명여자대학교 경영학부 교수
관심분야 소셜 미디어, 전자상거래, e-비즈니스, 정보시스템 관리