스마트카드 전자거래 빅데이터를 이용한 서울시 지하철 이동패턴 분석

Discovery of Travel Patterns in Seoul Metropolitan Subway Using Big Data of Smart Card Transaction Systems

김관호(Kwanho Kim)*, 오규협(Kyuhyup Oh)**, 이영규(Yeong Kyu Lee)***,
정재윤(Jae-Yoon Jung)****
{kwanhokim, k8383}@khu.ac.kr, magpie@smrt.co.kr, jyjung@khu.ac.kr

초 록

사람들의 이동흐름과 지리적으로 인접되어 있으면서 이동관점에서 같은 역할을 수행하는 Zone의 파악은 도시개발 및 이동편의성 개선 등을 위한 중요한 정보로 활용된다. 그러나, 기존의 연구는 특정 지점간의 이동과 Zone 발견을 개별적으로 수행하여, 거시적 관점에서 사람들의 이동을 이해하는 데 한계가 존재한다. 따라서, 본 연구에서는 스마트카드 전자거래 빅데이터로부터 Zone들을 발견하고 동시에 Zone들간의 관계를 설명하는 이동패턴 분석기법을 제안한다. 또한, 이동패턴을 효과적으로 평가하는 지표를 제안하여 이동패턴을 정량적으로 평가한다. 제안된 분석기법을 이용하여 서울시 지하철에서 수집된 실 데이터를 분석하여 서울시에서의 이동패턴을 밝혀내고 시각화하였다.

1. 서론

오늘날 도시환경에서 지하철은 사람들의 활동에 매우 중요한 역할을 수행하고 있다. 지하철 승하차 이동 데이터로부터 지리적으로 인접되어 있으면서 이동관점에서 동일한 역할을 수행하는 Zone들을 발견하고 이들 간의 관계를 파악하는 것은 도시개발 계획 수립 및 이동편의성 개선을 위한 중요한 정보로 인식되고 있다[1]. 여기서 Zone은 지리적으로 인접한 지역을 의미한다[2]. 본 논문에서도 유사한 의미로 인접한 지하철 역들의 집합을 의미하기 위하여 Zone이라는 용어를 사용한다.

본 연구에서는 스마트카드 전자거래 빅데이터로부터 Zone들을 발견하고 두 Zone들

이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단 기초연구사업의 지원을 받아 수행된 연구임(No. 2012003505).
* 경희대학교 산업경영공학과 박사후과정
** 경희대학교 산업경영공학과 박사과정
** 서울도시철도공사 영업계획팀
*** 경희대학교 산업경영공학과 조교수, 교신저자

간의 연관성을 나타내는 이동패턴(MZP: Movement Pattern between Zones) 분석기법을 제안한다. 구체적으로, 상향식 접근법(Bottom-Up Approach)을 적용하여 데이터로부터 실생활이 반영되는 의미 있는 이동패턴을 찾는 데 연구의 주안점을 둔다. 수집된 이동 데이터로부터 이동패턴을 추출하기 위한 병합적 군집화 기법(Agglomerative Clustering Method)을 개발하고, 설명률과 의존도에 의해 이동패턴을 정량적으로 평가하는 지표를 제안하였다.

제시된 분석기법의 효과는 서울시 지하철 5호선-8호선에서 수집된 승하차 이동 데이터를 이용하여 입증하였다. 나아가, 도출된 이동패턴으로부터 서울시에서 나타나는 주요한 이동특성을 분석하고 시각화하였다.

2. 관련연구

도시환경에서 지리적 이동패턴의 분석은 도시개발, 인구이동, 교통 등 다양한 관점에서 매우 중요한 주제로 인식되어왔다. 그럼에도 불구하고, 방대한 데이터의 수집과 분석기법의 한계로 인해서 그 동안 주로 설문조사를 통한 제한된 범위의 연구들이 수행되어왔다[3].

하지만 근래에 급속히 보급된 스마트카드로 인해 대량의 이동정보가 실시간으로 축적될 수 있게 되었고, 이를 이용한 자동화된 분석기법이 제시되었다[4]. <표 1>은 최근에 연구된 스마트카드 데이터를 이용한 이동패턴 분석연구를 나타낸다.

기존 연구에서는 지점간의 이동흐름과 Zone 발견을 개별적으로 접근하여, 거시적인 관점에서의 이동현상을 설명하는 데 사용될 수 없는 한계가 존재한다. 이와 달리, 본 연구에서는 Zone 발견과 동시에 이들간의 관계를 분석하는 데 목적을 두어 기존의 연구와는 차별화된다고 할 수 있다.

<표 1> 대중교통 스마트카드 데이터를 이용한 이동패턴 분석연구

연구내용	연구결과
지점간의 이동분석	[5], [6]
지점간의 이동예측	[7]
지점간의 관계분석	[8]
Zone 발견 및 분석	[2], [9]

3. 이동패턴분석

3.1 승하차 데이터 속성

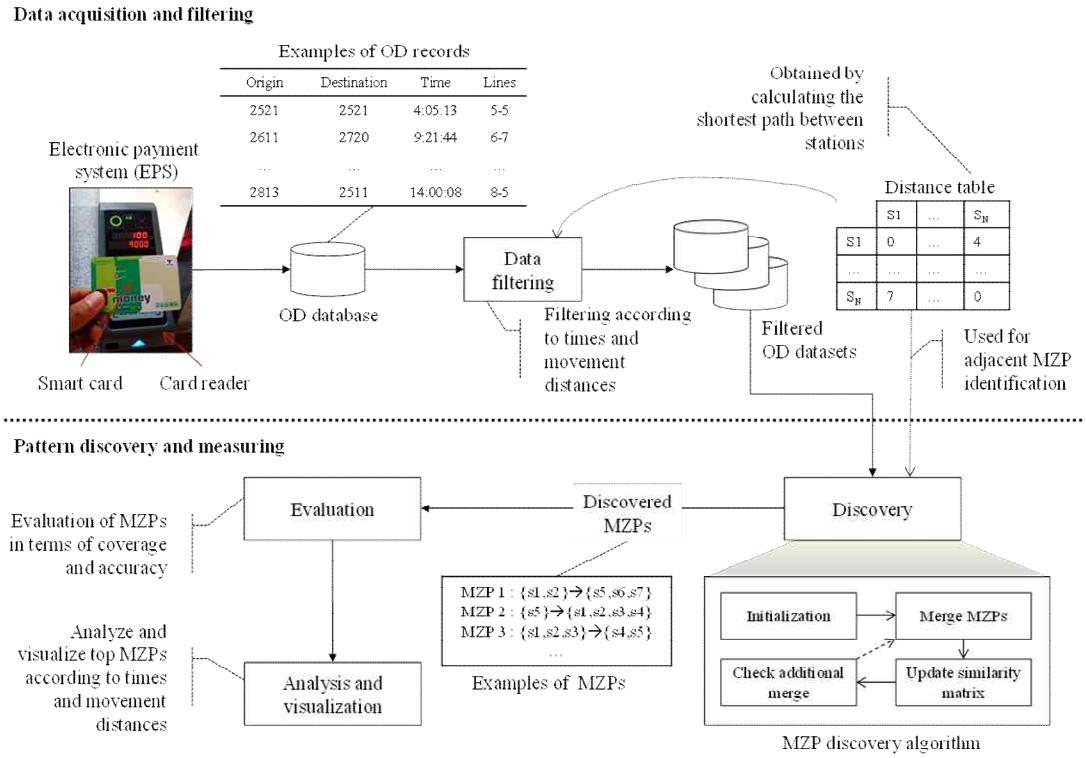
스마트카드 데이터는 수집방법과 활용목적에 따라 다양한 속성을 갖는다. 하지만 일부 속성은 특정 시스템에 종속적이어서 다른 환경에서 활용하는 데 제한적이다. 따라서, 본 연구에서는 <표 2>에서 제시된 네 가지의 기본적 승하차 정보만을 활용하여 제시된 모델의 적용범위를 최대화하고자 하였다.

<표 2> 사용된 승하차 데이터의 속성

속성명	설명 (타입)
Origin station	출발역 번호 (텍스트)
Destination station	도착역 번호 (텍스트)
Time	승하차 시간 (날짜, 시간)
Line	출발 및 도착역의 호선정보

3.2 이동패턴분석 프레임워크

본 연구에서의 이동패턴분석은 <그림 1>와 같이 크게 두 부분으로 구분될 수 있다.



<그림 1> 이동패턴분석 프레임워크

먼저, 데이터 수집 및 필터링 단계에서는 지하철의 스마트카드 시스템에서 기록된 승하차 이동 데이터를 수집하고, 이동거리 및 시간에 따라 분할하여 분석용 데이터베이스를 구축한다. 또한, 지하철 네트워크를 바탕으로 역간의 인접여부 및 최단거리를 계산하여 향후 이동패턴 분석에 활용할 수 있도록 한다.

다음으로, 이동패턴분석 및 평가부분에서는 이동패턴분석 알고리즘을 이용하여 데이터로부터 이동패턴을 분석한 후, 이들을 평가하여 이동패턴들을 정량적으로 비교 평가한다.

3.3 이동패턴분석 알고리즘

본 연구에서는 개별 데이터로부터 전체적인 관점에서 의미 있는 이동패턴을 발견하기 위해서 병합적 군집화 기법[10]을 이용한다. 기존의 병합적 군집화 기법은 주어진 유사도 함수에 기반하여 주어진 데이터를 트리 형태의 군집화 수행에 목적이 있지만, 본 연구에서는 이동 데이터의 방향성과 병합 시 Zone에서의 평균 이동 수를 고려하여 병합하는 부분을 수정하여 이동패턴 분석에 적합하도록 수정하였다.

<그림 2>는 제안된 이동패턴발견 알고리즘을 나타낸다. 초기화 단계에서는 개별 이동 데이터를 각각 이동패턴으로 설정하여 고유한 이동 수만큼의 이동패턴을 둔다. 다음으로, 이동패턴병합 단계에서는 모든 가능한 이동패턴의 쌍 중에서 가장 높은 Zone간 평균이동 수를 보이는 이동패턴의 쌍을 하나의 이동패턴으로 병합한다. 이동패턴병합은 더 이상 병합할 이동패턴이 없을 때까지 반복된다.

따라서, 이동패턴병합이 반복될수록 점차 이동패턴의 수가 줄어들게 되며, 알고리즘이 종료된 후 남은 이동패턴을 최종 이동패

1: **Initialize**
2: Set each station as a zone itself.
3: Build T MZPs each of which represents a distinct movement between zones.
4: Calculate the joint average frequency matrix $C^{(1)}$ based on $\rho_{i,j}^{(1)}$ for $i, j = 1, \ldots, T$.
5: **Repeat**
6: Merge p_i with p_j, $i, j = 1, \ldots, T-k$, $i \neq j$, if $\rho_{i,j}^{(k)}$, is the largest value in $C^{(k)}$.
7: Update the joint average frequency matrix $C^{(k)}$ to be $C^{(k+1)}$.
8: **Until** There is no MZP to be merged or the highest value of $\rho_{i,j}^{(k)}$, for $i, j = 1, \ldots, T-k$, is less than θ.
9: **Return** The remaining $(T-k)$ MZPs.

<그림 2> 이동패턴분석 알고리즘

턴으로 삼는다. 이동패턴이 병합되면서 Zone의 범위가 넓어지기 때문에 남은 이동패턴들은 기존의 것들과 비교해서 우수한 설명률을 보이게 된다. 또한, Zone간 평균 이동을 고려하기 때문에 Zone간의 의존도 측면에서도 양호한 이동패턴을 추출할 수 있게 된다.

제시된 알고리즘에서 $\rho_{i,j}$ 는 i-번째와 j-번째 이동패턴을 병합한 후의 Zone간 평균 이동 수를 나타낸다. 만일 두 이동패턴이 병합된 후의 Zone에 인접하지 않은 역이 존재할 경우 $\rho_{i,j} = 0$, 그렇지 않은 경우에는 다음과 같이 정의된다.

$$\rho_{i,j} = \frac{\#(O_i \cup O_j \rightarrow D_i \cup D_j)}{|O_i \cup O_j| \cdot |D_i \cup D_j|}$$

여기서 $\#(\cdot)$ 는 주어진 이동패턴이 설명할 수 있는 이동 수를 나타낸다.

4. 이동패턴평가

본 연구에서는 이동패턴을 위해 다음과 같은 세 평가지표를 제시한다. 제시되는 각각 지표들은 장바구니 분석에서 규칙의 평가를 위해 제안된 지표들인 Support, Lift, Cosine를 기반으로 한다[11]. 기존의 지표들은 동시에 발생하는 사건들에 대해 평가하는데 목적을 두고 있어서, 이를 Zone 내의 임의의 역에서 발생하는 이동을 고려하여 평가할 수 있도록 수정하였다.

첫째, 이동패턴이 얼마나 많은 이동을 설명할 수 있는지를 나타내는 지표(v-value)를 제시한다. 구체적으로, Zone O_i 중 한 역에서 Zone D_i 중 한 역으로의 이동을 나타내는 이동패턴 $O_i \rightarrow D_i$ 의 v-value은 다음과 같이 정의된다.

$$v(O_i \rightarrow D_i) = \Pr(O_i \rightarrow D_i) = \frac{f_{i,i}}{M}$$

여기서, $f_{i,i}$ 은 O_i 중의 한 역에서 탑승하고 D_i 중의 한 역에서 하차한 이동 수를 나타내며, M 은 관측된 총 이동 수를 나타낸다.

둘째, 두 Zone들이 얼마나 큰 상호 종속성을 가지고 있는지를 고려하는 지표(a-value)를 제시하고, 이동패턴 $O_i \rightarrow D_i$ 의 a-value는 다음과 같이 계산된다.

$$a(O_i \rightarrow D_i) = \frac{\Pr(O_i \mid D_i)}{\Pr(O_i)} = \frac{\Pr(D_i \mid O_i)}{\Pr(D_i)} = \frac{Mf_{i,i}}{f_{i,*} f_{*,i}}$$

여기서, $f_{i,*}$ 과 $f_{*,i}$ 는 각각 O_i 중의 한 역에서 탑승한 모든 이동 수와 D_i 중의 한 역에서 하차한 모든 이동 수를 의미한다.

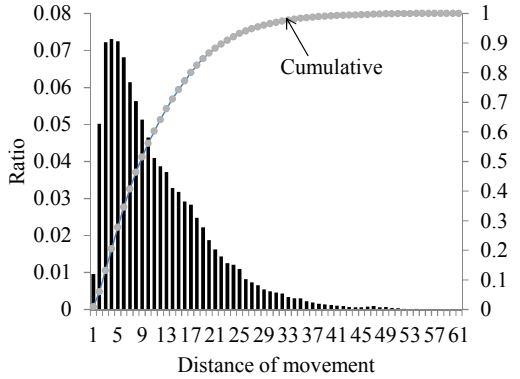
마지막으로, 앞서 제시된 두 지표를 같이 고려하는 복합지표(c-value)를 제시한다. 이동패턴 $O_i \rightarrow D_i$ 의 c-value는 다음 식을 통해 얻어진다.

$$c(O_i \to D_i) = \sqrt{v(O_i \to D_i)a(O_i \to D_i)} = \frac{f_{i,i}}{\sqrt{f_{i,*}f_{*,i}}}$$

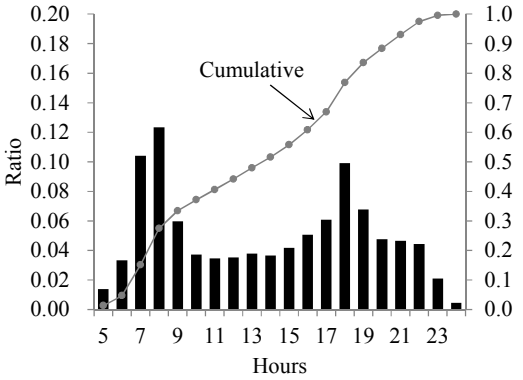
5. 실험결과

제시된 기법의 효과를 검증하기 위해서 서울도시철도공사에서 2012년 6월 18일부터 22일까지 5호선-8호선에서 수집된 총 5,405,736건의 승하차 이동 데이터 사용하였다. 승하차가 기록된 시간은 05시부터 23시까지이며, 총 역의 수는 148개였다. 수집된 데이터는 5, 6, 7, 8호선이 각각 32%, 25%, 27%, 14%의 분포를 보였다.

<그림 3>의 (a)는 이동거리에 따른 데이터 분포를 나타내며, 이동 역 개수의 평균은 10.04, 중앙값은 8, 최빈값은 4로 나타났다. 가장 빈번하게 관측된 역간의 이동은 "광명사거리역"에서 "가산디지털단지역"이였으며, 전체 중 0.48%의 이동이 이에 해당되었다. 또한, 전체 21,904개의 출발역과 도착역의 쌍들 중 97개의 쌍이 전체 이동 데이터 중 10%가 넘는 이동에 해당되며, 2,265개의 쌍들은 이동이 관측되지 않았다.



(a) 이동거리에 따른 데이터 분포



(b) 시간에 따른 데이터 분포

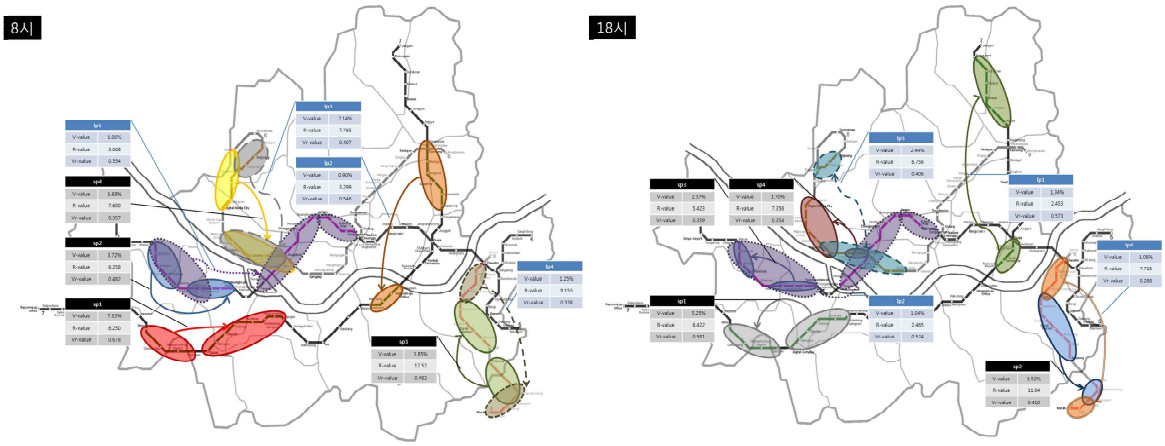
<그림 3> 수집된 데이터 분포

이는 다수의 이동이 소수의 역에 집중되고, 다수의 역은 소수의 이동만이 발생함을 의미한다.

<그림 3>의 (b)는 시간에 따른 이동 데이터 분포를 나타내며, 출근 시간대인 8시

<표 3> 서울도시철도 데이터를 통해 분석된 주요 이동패턴

Rank	Origin Zone	Destination Zone	Measures		
			v-value (%)	*a*-value	*c*-value
1	철산 - 온수	상도 - 가산디지털	1.552	5.659	0.296
2	장승배기-가산디지털	철산 – 온수	1.280	6.122	0.280
3	발산 – 까치산	방화 – 송정	0.794	7.193	0.239
4	방화 – 송정	발산 – 까치산	0.782	7.261	0.238
5	암사 – 몽촌토성	잠실 – 가락시정	0.580	7.934	0.215
6	신정 – 영등포구청	발산 – 까치산	0.939	4.847	0.213
7	발산 – 까치산	신청 – 오목교	0.785	4.873	0.196
8	마포구청 – 합정	광흥창 – 공덕	0.402	8.626	0.186
9	광흥창 – 공덕	마포구청 – 합정	0.388	8.803	0.185
10	태릉입구 – 용마산	청담 – 논현	1.336	2.469	0.182



(a) 출근 시간대(8시)의 주요 이동패턴 (b) 퇴근 시간대(18시)의 주요 이동패턴

<그림 4> 제안된 기법을 통해 분석된 서울시의 주요 이동패턴의 시각화

와 퇴근 시간대인 18시에 가장 높은 이동이 발생함을 알 수 있다. 해당 출퇴근 시간에 전체 이동 중 22%가 발생하였다.

주요 이동패턴을 살펴보기 위해 수집된 데이터로부터 <그림 2>에 제시된 알고리즘을 이용하여 이동패턴을 추출하였다. 밝혀진 이동패턴들은 제안된 세 지표로 평가되었다. <표 3>은 복합지표를 기준으로 상위 10개의 이동평균을 나타내고 있다. 복합지표 관점에서 가장 뚜렷한 패턴을 보이는 Zone들은 7호선의 “철산역”-“온수역” 구간과 “상도역”-“가산디지털단지역” 구간이었다. 해당 이동패턴은 전체의 1.55%에 해당되는 이동을 설명하며, 동시에 매우 높은 의존성을 보였다. 따라서, 각각의 Zone의 역들은 이동관점에서 단일기능을 수행하고, 두 Zone은 유입과 유출 관점에서 서로 큰 연관성을 가지고 있다고 볼 수 있겠다.

또한, 3번째와 4번째 이동패턴들은 정확히 같은 Zone들을 가지고 서로 방향이 반대임을 알 수 있다. 이는 출근 및 퇴근과 같이 시간에 따라 이동방향이 바뀌는 현상 때문으로 판단되며, 해당 Zone들은 양방향 모두 높은 의존성을 보이고 있다. 따라서, 두 Zone들은 각각 서로에 대한 유출지역과 유입지역의 역할을 수행하고 있는 것으로 볼 수 있겠다.

마지막으로, 시간대별 이동패턴들의 특징을 살펴보기 위해 <그림 4>와 같이 출근 시간대(8시)와 퇴근 시간대(18시)에 대해서 어떻게 이동패턴이 변화하는지를 시각화하였다. <그림 4>의 (a)는 복합지표 기준으로 8시에 보여지는 상위 5개의 이동패턴들을 나타낸다. 출근 시간대에는 외곽지역(거주지역)에서 중심지역(상업지역)으로의 이동패턴을 확인할 수 있다. 또한 <그림 4>의 (b)는 18시에 보여지는 상위 5개의 이동패턴을 나타낸다. 퇴근 시간대에는 출근 시간대와 유사한 Zone들에서 방향이 반대인 이동패턴이 관측됨을 알 수 있다.

5. 결론

본 논문에서는 스마트카드 빅데이터를 이용하여 이동패턴을 추출하여, 지리적으로 유사하면서도 동일한 기능을 수행하는 Zone을 발견하고 이들간의 연관성을 파악

하고자 하였다. 또한, 추출된 이동패턴을 정량적으로 평가하기 위한 지표를 제안하였다.

연구의 결과는 지하철뿐만 아니라 버스, 택시 등의 대중교통에서 발생하는 대용량의 이동 데이터를 바탕으로 보다 다양한 이동 분석을 가능하게 할 것이다. 나아가, 제시된 분석결과는 도시계획, 대중교통 서비스 향상, 대체 이동수단 보완 등에 활용될 수 있을 것으로 기대된다.

참고문헌

[1] Yuan, J., Zheng, Y., Xie, X., "Discovering regions of different functions in a city using human mobility and POIs", Proceedings of the18th ACM SIGKDD International Conference on Discovery and Data Mining, Vol. 12, pp.186-194, 2013.

[2] Fusco, G., Caglioni, M., "Hierarchical Clustering through spatial interaction data. The case of commuting flows in south-eastern France", Lecture Notes in Computer Science, Vol. 6782, pp.135-151, 2011.

[3] Bagchi, M., White, P.R., "The potential of public transport smart card", Transport Policy, pp.464-474, 2005.

[4] Blythe, P., "Improving public transport ticketing through smart cards", Proceedings of the Institute of Civil Engineers, Municipal Engineer, Vol. 157, pp. 47-54, 2004.

[5] Jang, W., "Travel time and transfer analysis using transit smart card data", Journal of the Transportation Research Board, Vol. 2144, pp.142-149, 2010.

[6] Srinivasan, S., Ferreira, J., "Travel behavior at the house level: understanding linkages with residential choice", Transportation Research Part D, Vol. 7, pp.225-242, 2003.

[7] Park, J.Y., Kim, D.J., "The potential of using the smart card data to define the use of public transit in Seoul", Journal of the Transportation Research Board, Vol. 2063, pp.3-9, 2008.

[8] Trépanier, M., Morency, C., Agard, B., "Calculation of transit performance measures using smartcard data", Journal of Public Transportation, Vol. 12, No. 1, pp. 79-96, 2009.

[9] Konjar, M., Lisec, A., Drobne, S., "Method for delineation of functional regions using data on commuters", Proceedings of the 13-th AGILE International Conference on Geographic Information Science, Portugal, 2010.

[10] Day, W., Edelsbrunner, H., "Efficient algorithms for agglomerative hierarchical clustering methods", Journal of Classification, Vol. 1, Iss. 1, pp.7-24, 1984.

[11] He, B., Ding, Y., Yan, E., "Mining patterns of author orders in scientific publications", Journal of Informetrics, Vol. 6, No. 3, pp.359-367, 2012.