마코프모형의 계량정보학적 응용연구

A Study on Markov Chains Applied to Bibliometric

문경화, 중앙대학교 문헌정보학과

Moon Kyung Hwa, Dept, of LIS, Chung Ang Univ.

계량정보학 연구영역의 하나인 운영연구(Operation Research, OR)중, 미래예측이라는 목적을 가지고 있는 마코프모형(Markov chains)의 통계기법을 활용한 두가지 실험사례를 살펴보고, 최근의 연구경향을 분석함으로써, 도서관 시스템 운영과 설계에 마코프모형을 응용할 수 있는 네가지 방안을 제시하였다. 계량정보학의 한분야로 적용되고 있는 마코프모형에 관한 국내연구가 활발하지 못한 상태이므로, 국내 계량정보학에서의 마코프모형 연구의 필요성과 활성화를 제안하였다.

1. 서 론

학문의 계량적 분석연구는 학문의 이론과 현실간의 괴리를 제거하기 위한 목적으로, 인문사회과학 여타 분야에서와 마찬가지로 계량정보학에서도 그 중용성이 강조, 응용되고 있다. 국내 계량정보학 연구영역에 있어서 수량적 접근방법을 통한 확률이론인 마코프모형의 연구를 통해서, 도서관 시스템의 여러부문에서도 효과적으로 응용할 수 있는 가능성이 있음을 인식하고, 이에 연구필요성을 가지고 본 연구를 시작하였다. 본 연구에서는 계량정보학의 연구영역중 시스템의 운영, 계획 및 설계에 관한 문제를 계량적, 체계적으로 분석, 해결해 보고자 하는 운영연구(OR)의 연구기법인 수리계획법, 네트워크모형,대기행렬이론, 마코프모형, 모의실험, 결정이론가운데에서 마코프모형(Markov chains)을 중심으로 도서관 시스템의 운영, 설계에 적용하기 위한 방안을 모색해 보고자 하였다.

2 이론적배경

2.1 마코프모형의 특성

이러한 계량정보학에서의 마콥체인을 설명하기 위해서는 먼저 확률과정을 이해하여야 하는데, 일반적으로 시간의 연속성 여부에 따라 이산-시확률과정(discrete-time stochastic process)과연속-시확률과정(continuous-time stochastic process)으로 구분된다. 이산-시확률과정은 확률과정에서 보조변수인 시간t가 일정한 값을 취하며 변화하는 유한적 경우이고, 연속-시확률과정은 t가 연속적으로 변화하는 무한적 경우를 말한다. 마코프모형은 이러한 개념의 이산-시 확률과정의 특별한 경우로서 n+1번째 단계 상태확률이 n번째 단계 상태확률의 영향만 받고, 그 이전 단계의 상태확률의 영향을 받지 않는다는 마콥성질을 갖는다. 마코프모형으로 기술하기 위한 세가지 특성을 보면 다음과 같다.

첫째, 그과정은 이산-시 확률과정이어야 하고 둘째, 상태공간이 유한집합이어야 하며 셋째, 그과정은 각 상태의 예측된 확률은 바로 이전 단계의 상태에만 의존하여야 한다는 것이다. 이러한 마코프모형의 특성을 설명하기 위해서 연구기관 내에서 연구자들의 전직과정을 예를 들어보면 다음과 같다.

[표1] 최근 4년간의 연구자들 이동과정

a	b	c
A_1	C_1	B
C_2	D_1	D_2
A_2	B_2	D_3
A_3	B_3	D_4
C_3	B_4	A_4
C_4		

[표2] 연구자들의 연구기관 이동추이

M^1 =

	a	b	c	계
a	4	0	1	5
b	1	2	1	4
c	0	1	2	3

세 개의 연구기관 a,b,c가 있고 여기서 근무하는 네명의 연구자 A,B,C,D가 있다고 가정한다. 각 연구자는 어느 한 연구기관에 고용되며, 연초마다 다른기관으로 옮기거나 그대로 머물 수 있다고 할 때, 연구자들이 선택할 수 있는 연구기관을 "상태", 상태의 집합을 "상태공간"이라고 할 수 있다. 여기에서 보면, 우선 1년간격으로 연구기관을 옮길수 있고, 그대로 머무를 수 있기 때문에 이산-시 확률과정이며, 두번째

로 상태공간이, a,b,c로 한정되어 있는 유한집합이며, 세번째 현재의 연구성과에 의해서 연구기관을 옮기는데 영향을 받게 될 것이므로 각 연구자의 다음상태는 현상태에 의해서 결정된다고 볼 수 있는 것이다.

2.2 초기상태확률과 추이확률

이상의 세가지 특성을 충족시키는 것 이외에 각 상태의 초기확률과 추이확률이 필요하다. 실험을 시작할때, 각 연구기관에 대한 연구자들의 비율에 대한 합리적인 가정을 해야 하는데, 각 상태의 초기확률은 첫해에 각 연구기관에 고용된 연구자수를 총연구자수로 나눈 값으로 얻어진다. 각 연구기관에 대한 연구자들의 비율에 대한 초기확률이 구해지고나면, 한 연구자가 현 연구기관에서 세가지 연구기관중 어느하나로 이동할 수 있는 가능성의 추이확률이 필요하다. 추이확률은 현재상태 i에서 다음상태 j로 이동될 확률로 일반적으로 P_{ij} 로 나타낸다. P_{ab} 의 조건부 추이확률은 상태 a에서 상태 b로 이동할 확률이기 때문에, a 연구기관에서 b 연구기관으로 이동하는 경우를 계산함으로써 얻어질 수 있다.
즉 [표1]에서 보면 첫 해에는 a 연구기관에 연구자 A 한명이, b 연구기관에 연구자 C 와 D 두명이, 그리고 c 연구기관에 연구자 c 한명이 고용되어 있었다. 따라서 기관 a, b, c 에 대한 초기확률은 아래의 [표3]과 같게 된다.

[표3] 초기상태확률

a	b	c
0.25 (1/4)	0.50 (2/4)	0.25 (1/4)

[표4] 추이확률행렬

M^1 =

a	0.80	0	0.20
b	0.25	0.50	0.25
c	0	0.33	0.67

마찬가지로 추이확률은 [표4]와 같이 나타날 수 있는데 제1행을 보면 a라는 상태에서 각각 a, b, c 라는 상태로 옮겨갈 확률을 보여주고 있다. 상태 a 에서 한번 이동할 때 a, b, c중 어느 한 상태로 옮겨가게 되므로 반드시 그 행렬에서의 확률합은 1이 성립된다. 즉 $P_{aa} + P_{ab} + P_{ac} = 1$이 성립된다. 이 경우에 P_{aa} 의 조건확률은 상태 a에서 상태 a로 이동할 확률이기 때문에 이 경우는 $A_1 \rightarrow A_2$. $A_2 \rightarrow A_3$. $C_2 \rightarrow C_3$. $C_3 \rightarrow C_4$로 총 4번이며 a 연구기관에서 b연구기관으로 옮겨간 사람은 아무도없고 a 연구기관에서 c연구기관으로 옮긴경우는 $A_3 \rightarrow A_4$ 로 단 한번 있다. 따라서 상태 a 에서 a, b, c로의 추이확률은 각각 4/5(=0.80), 0, 1/5(=0.20)이 된다.

2.3 미래예측과 고정확률

마코프모형의 중요한 목적은 시스템의 미래의 행위를 예측할 수 있다는 것이며, 그 결과는 추이확률과 현 단계의 상태확률이 결합되어 구해지는 다음단계의 상태확률과 추이확률을 n제곱승 하여 구할 수 있는 시스템의 장기적인 안전상태의 확률을 나타내고 있다. 예에서 다음기간의 각 연구기관에 고용될 연구자들의 비율 즉 상태확률을 예측하고자 하는데, 이 예측은 초기확률과 추이확률을 곱하여 구할 수 있다. 다음기간의 상태확률이 계산된후 그 다음기간의 상태확률의 계산도 가능한 것이다. 두 기간 후의 상태확률은 다음 기간의 상태확률과 추이확률을 곱해서 구해지며 [표5]와 같다.

[표5] 미래예측의 추이확률(2승)

M^2 =

a	.640	.067	.293
b	.325	.333	.342
c	.083	.389	.528

마코프모형 이론에 의하면 추이확률 P가 정규마코프모형이면, 여러단계를 거친후 각행에 있는 확률값들의 분포가 똑같은 고정확률행렬 L을 얻게 되는데, 일단 고정확률행렬 L에 이르면 행렬의 확률값들은 시간에 따라 더 이상 변화하지 않는다. 이경우는 2의 5승 단계를 거친후 고정확률행렬에 이르게 된다. 고정확률행렬은 몇 년후에 연구자들의 33%, 27%, 40%가 연구기관 a,b,c에서 각각 근무하게 될 수 있음을 예측하고 있다. 즉 초기확률과 고정확률을 비교하면 c 연구기관으로 연구자들이 몰릴 것을 예측할수 있는 것이다.이는 c연구기관의 고정확률이 40%로 초기확률인 25%보다 훨씬 높기 때문이다.[표6] 결국 마코프모형은 가상의 연구기관실험에서 수년후에 세 연구기관의 합리적규모를 예측할 수 있는 패턴을 지시해 주는 것이다.

[표6] 고정확률L(5승)

M^5 =

a	.3335	.2662	.3996
b	.3332	.2664	.3997
c	.3328	.2667	.3997
	0.33	0.27	0.40

2.4은닉마코프모형(HiddenMarkovChains)

은닉 마코프모형(HMM: Hidden Markov Model)은 이중적으로 결합된 stochastic process로 구성된 확률적 함수이다. HMM의 내부에 존재하는 것으로 가정되는 Markov chain은 유한한 개수의 상태와 각 상태와 결부되어 있는 난수 함수들의 집합을 가지고 있다. 각각의 이산시간에서 프로세스는 어떤 상태에 있고, 그 상태를 결정하는 난수 함수로부터 한 출력벡터가 관측된다고 가정한다. 그리고 나서 내부의 마코프모형은 전이확률행렬에 따라 다

음 상태로 전이하게 된다. 따라서 관측자의 입장에서는 오직 상태로부터의 출력벡터만을 관측할 수 있을뿐이며, 내부의 상태는 관측할 수 없는 상황이다. HMM은 한 상태에서 관측가능한 벡터들에 따라 이산분포, 준연속분포, 연속분포등으로 분류된다. 즉 은닉 마코프모형은 지금까지의 이산-시(discret-time) 개념의 이론에서 달리 연속적인(continuous)개념의 이론을 적용한 것이다. 이러한 은닉 마코프모형은 군집개념을 구성하는데 클러스팅으로서 많이 활용되고 있다.

3. 실험사례

먼저 민속음악 주제에 관한 저자데이터 비교실험을 보면, 10년동안의 2018명의 저자가 발표한 3302건의 발행물을 대상으 로 9개의 하부주제를 설정하였다. 각저자가 첫해에 발표한 논문은 9개주제 a에서 i까지 각각 52, 52, 148, 409, 614, 194, 175, 35, 339개의 논문들이 초기확률로서 나타났다. 그다음은 한주제에서 다른주제에로의 이동을 보여주는 추이행렬을 통해서 다른주제로 이동할 확률인 추이확률을 구할 수 있다. 이러한 초기확률과 추이확률을 비교하여 결과를 보면 1977년부터 1978년동안 유럽음악은 예측되었던 바처럼 지배적인 우의를 차지했다. 아시아음악도 같은 기간중에 높은 발행물이 발표되었다. 북미음악은 1979년에서 1980년에 감소경향을 보여주고 있으며, 아시아와 유럽민속음악에서 발행물이 집중되는 것을 발견할 수 있었다. 이러한 인문분야 주제에서의 실험은 한 주제영역에 대한 연구에서 저자들의 이동유형을 예측하기 위해서 마코프체인의 특성을 사용할 수 있다는 것을 보여준다. 이런 과정의 활용은 많은 영역에서 광범위한 영향을 미쳤다.

두 번째 실험은 한국화학자들의 7개 하위주제간 이동실험인데 두가지 실험중 시스템의 미래의 행위를 예측하는 기법인 마코프모형을 이용하여 학문의 발전과정을 기술,예측할 수 있다는 가설에 대한 첫실험만 보면, 우선<학술총람 제33집>에서 1967년에서 1973년동안 2편이상의 논문을 쓴 313명의 화학자들이 저술한 총논문수 968편을 분석하여 각 하위주제의 초기상태확률, 추이확률 및 고정확률을 구하였다. 고정확률에 의해서 예측된 각 하위주제에 대한 연구자들의 비율이 타당성이 있는지를 검증하기 위해서 313명중에서 1974년부터 1983년동안 1편이상의 논문을 쓴 290명을 선정하여 이들이 쓴 총논문수 1,773편을 7개의 하위주제별로 집계하여 분석한 결과인 초기확률을 고정확률과 비교해 본결과, 예측된대로 무기,분석 유기화학에 관한 논문을 발표한 연구자들의 비율이 낮아졌으며, 물리,생,고분자화학에서는 그들의 비율이 높아지고 있었다. 그러나 공업화학만은 연구자들의 비율이 조금 높아지리라는 예측과는 달리 다소 낮아지고 있었다. 결론적으로 7년동안 화학자들의 하위주제간의 추이과정을 분석하여 얻은 각 하위주제의 고정확률은 앞으로의 연구자들의 확률분포를 대체적으로 예측해 주고 있다고 말할 수 있었다.

4.계량정보학에서의 마코프모형응용

Zunde와 Slamecka은 마코프모형을 응용한 학문개발처리 모형을 정립한 바 있으며, 고프만은 문헌의 동적과정 연구에서 마코프모형을 응용하였고, 기호를 사용하여 하위주제에서 저자들의 변동상황을 설명하고 예측하기 위해 마코프모형을 사용하였다. 그는 주혈흡중병과 수두세포 주제분야에서 저자를 분석하기도 하였다. 최근에 들어서 마코프모형이론은 물리학과 사회과학 그리고 병리학, 기술과학, 경영학에서 많이 응용되고 있음을 볼 수 있는데 이러한 마코프모형을 계량정보학에 응용할 수 있는 몇가지 제안을 한다.

4.1 장서개발 및 수서정책

도서관에서의 장서개발 특히 수서정책에 있어서 마코프모형을 이용하여 수서시스템의 미래행위를 예측하고, 효과적인 운영을 하는데 응용할 수 있다. 저자의 저술실태를 조사함으로써, 장서개발에 있어서 강화,유지, 혹은 쇠퇴될 학문영역에 대한 수서기준의 과학적이고 보다 객관적인 근거자료로써 활용할 수 있다. 최근 10년동안의 9개 학문분야들에 대한 저술실태가 변화하는 것을 조사한다고 하면, 우선 상태를 저술할 학문분야들로 보고, 초기확률은 첫해에 9개 학문분야에서 발표된 저술건수들 각각을 그해 전체 저술건수로 나눈값이 된다. 이를 통해서 주제영역별 변화추이값인 추이확률을 구하고, 패키지를 통해서 안정된 확률분포를 나타내는 고정확률을 구할 수 있다. 여기서 초기확률과 고정확률을 비교해보면, 가까운 미래에 발전할 학문분야와 쇠퇴될 영역을 예측할 수 있다. 즉, 도서관 이용자의 현재와 미래 연구관심영역의 경향을 예측하여, 이에 대응한 도서관 장서를 구성하고 보다 적합한 서비스를 준비할 수 있다는데 그 응용효과가 있다.

4.2 저자이동연구

마코프모형 기법은 특별히 관련성이 있는 저자들그룹 사이에서 그들의 주요연구분야가 변화하는 것에 의해서 특정저자의 연구분야 이동연구에 응용될 수 있다. 마찬가지로, 특정 저널그룹 사이에서 발표경향은 이러한 저널들 사이에서 인용의 이동추이에 따르게 될 것이다. 즉, 이러한 저자 및 인용 이동추이를 통해서 미래의 연구경향을 예측할 수 있는데 마코프모형이 응용될 수 있다.

4.3 대학간 교수이동

대학에 소속된 교수들이 일정기간동안, 일정대학들 사이에서 이동하는 추이를 또한 마코프모형을 통해서 예측할 수 있다. 최근 십년간 a대학에서 f 대학에 소속하여 근무하는 교수들

이 그들 대학에 매년 옮겨 갈 수 있거나 그대로 머물 수 있다고 가정하면, 상태롤 이동할 수 있는 대학으로 볼 수 있다. 마코프모형의 첫번째 조건인 이산-시 확률과정을 충족시키고, 옮겨 갈 수 있는 상태공간을 6개 대학 (a, b, c, d, e, f)으로 한정함으로써, 마코프모형의 두번째 조건을 충족시킨다.또한 현재의 연구성과나 관심에 따라서 옮겨 가게 될 것이므로 세번째 조건까지 모두 충족된다. 여기서 초기확률은 각 대학에 소속된 첫해 교수수를 전체 교수수로 나누어 산출할 수 있으며, 각 교수들의 대학별 이동추이를 추적하여, 추이확률을 구할 수 있다. 이를 통해서 어느 전공분야의 교수들이 어느 특정대학으로 몰릴 가능성을 예측해볼 수 있게 될 것이며, 이를 통해서 대학별 성장학문의 특성화를 가늠해볼 수 있는 연구효과가 있는 것이다.

4.4 재정정책

마코프모형은 재정정책 수립에도 영향을 미칠 수 있는데, 하부주제에서 인용이나 저술 이동과정을 추적함으로써, 한정된 학문내에서 연구관심분야의 이동연구에 관심을 가지고, 그런 정보에 근간해서 학문적으로 성장 기대되는 분야와 퇴보하는 분야에 대한 기준을 가지고 재정정책을 세울 수 있다. 즉 인용이 많이 되는 영역의 자료구입 예산배정을 높게 잡고, 저술실태가 저조하거나 자주 인용되지 않고 있는 영역의 자료구입 예산배정을 낮게 혹은 삭감할 수 있도록 예측하여,도서관 시스템운영을 위한 보다 과학적이고 합리적인 예산수립을 세울 수 있다. 분야별 이용율 즉 대출율을 구하여 마찬가지로, 초기확률, 추이확률 그리고 고정확률을 구할 수 있을 것이며 이를 비교하여 예산배정을 위한 합리적이고 객관적인 예측으로 시스템 운영을 할 수 있는 마코프모형의 응용효과가 있다.

이상의 응용분야에서 이런 기법을 적용함에 있어서 추이행렬의 검증을 계산하는 과정이 상당히 많은 단계를 제곱승 하는 번거로움이 생기는데 이런과정은 MINITAB 패키지 등을 사용하여 보다 쉽고 간편하게 응용 할 수 있다. BASIC 해석기 또한 유용할 수 있으며, SAS나 SPSS 그리고 최근에는 GAUSS를 보편적으로 적용할 수 있다.

5. 계량정보학의 새로운 파라다임 -사이버메트릭스(Cybermetrics)

마코프모형은 계량정보학의 한 분야로서 적용되고 있으며, 이러한 계량정보학의 새로운 영역으로 사이버메트릭스(Cyermetrics)가 연구되고 있다. 기존의 계량정보학이 학술서지문헌을 대상으로 측정과 계량을 시도하였다면, 최근에는 인터넷 등의 폭발적 보급과 활용으로 인해 웹 등의 사이버상에서의 측량과 분석이 이루어지고 있는 새로운 계량정보학적 연구가 나타나고 있는 것이다. 앞으로 사이버메트릭스를 통하여 인터넷 접속 대기시간 및 서치엔진 검색결과 대기시간 등을 마코프모형으로 측정하여 향후의 발전모델을 정립하는데 유용할 수 있으리라 본다. 한예로, 1996년 JASIS에 발표된 Ray R. Larson의 Bibliometrics of the World Wide Web이라는 논문은 사이버스페이스에서의 지적구조를 분석하여 웹에 대한 계량서지학적 연구한 바 있는데, Web Crawler와 Altavista로 수집한 주제의 웹사이트에 대한 계량서지학을 시도한 연구로서, 웹문헌의 통계적 특성과 이들의 하이퍼텍스트링크와 자주 인용되는 웹문헌에 대한 특징을 분석하였다. 결론적으로 말해서 유사성을 가진 웹사이트의 군집화는 매우 합리적으로 나타났으며, 단일분야에 대한 이 연구방법을 응용한다면 다른 학술문헌 분야에도 적용할 수 있을 것이고 향후 계속 연구해야 할 과제이다.

6. 결론 및 향후연구

이상에서 마코프모형에 대한 기본적인 이론과 이모형을 활용한 두가지 실험사례를 통해서 계량정보학에서의 응용방안을 제안하였다. 그러나, 보다 복잡한 상황 즉 휴간이나 폐간저널에 대한 문제점이 있었으며, 마코프모형이 학문발전과정을 기술하고 예측할 수 있다는 가설을 검증하는 데 있어서 보다 충분한 실험데이터를 가지고 진행될 수 있어야 할 것이다. 본 연구에서는 마코프모형의 도서관 실무분야에 대한 응용 가능성을 제안하는 선에서 이루어졌으며 실제실험은 수행되지 않았다. 향후의 연구에서 구체적인 상태를 설정하여 심도깊은 실험을 계속하고자 한다.

참고문헌 및 URL

1. Introduction to mathematical statistics fourth editions by Robert V.Hogg & Allen T. Craig
2. Introduction to stochastic processes by Erhan cinlar p.106-143
3. Queuening systems by Leonard Kleinrock p.26-43
4. An Introduction to probability theory and its applications by William Feller p.321-357
5. Bibliometrics of the World Wide Web by Ray R. Larson JASIS 1996. no.33
6. A Quality filtering method for biomedical literature by Seung-Hee Sohn
7. Bibliometric application of Markov Chains by Miranda Lee Pao & Laurie McCreery Information Processing & Management vol.22. No.1 p.7-17
8. http://www.cindoc.csic.es/cybermetrics/links03.html
9. http://sci.hkbu.edu.hk/math/markov.html