미국 인구통계 데이터를 이용한 분산형 데이터마이닝 시스템 성능평가

김충곤°, 우정근, 김성국, 백성욱
세종대학교 컴퓨터공학부 지능형 미디어 연구실
forgom°@gmail.com, yakumo4@gmail.com, realpyromania@hanmail.net, sbaik@sejong.ac.kr

The evaluation of Distributed Data Mining System using USA census Database

Choong Gon Kim°, Jung Geun Woo, Kim Sung Guk, Sung Wook Baik
Intelligent Media Laboratory, Sejong University

요 약

본 논문에서는 분산형 환경에 적합한 새로운 의사결정나무 알고리즘을 제안하고 그 실용성을 확인하기 위해 분산형 데이터마이닝 시스템을 구현하였다. 그리고 본 논문에서 구현한 시스템을 평가하기 위해 데이터의 신뢰성이 높은 방대한 양의 미국의 인구통계 데이터(Census bureau database)를 사용하였다. 본 논문에서 구현한 시스템을 이용하여 신뢰성을 테스트하였고 그 결과가 다른 시스템의 알고리즘과 유사한 신뢰성을 나타내었다.

1. 서 론

데이터의 양이 급속도로 증가하는 현대 사회에서 대량의 데이터를 분석하기 위해 데이터마이닝작업을 할 때 계산의 효율성과 확장성은 매우 중요하다. 일반적으로 대량의 데이터들은 분산된 위치에 있는 각 기관이나 조직들의 데이터베이스 시스템에 위치하고 있고 일부 데이터들은 서로 다른 위치에 저장되어 있을지라도 매우 관련성이 높을 수 도 있다. 위와 같은 상황에서 데이터마이닝의 새로운 연구 분야는 각기 다른 장소에 있는 대량의 데이터를 함께 분석하여 유용한 지식정보를 제공할 수 있는 방법론을 제공하는데 초점이 맞춰지고 있다.

일반적으로 분산 환경에서 데이터마이닝을 하기 위한 첫 번째 방법은 분산된 데이터를 중앙에 모두 모은 후 중앙집중식 데이터마이닝으로 분석을 하는 것이다. 이 방법은 어떠한 방법보다도 정확한 분석결과를 얻을 수 있는 반면 많은 계산 비용과 데이터 이동 비용이 들게 된다는 단점이 있다. 또한 각 기관들의 중요하고 민감한 정보들을 서로 공유 한다는 것은 현실상 불가능한 일이다. 두 번째 방법은 분산되어 있는 데이터를 한곳으로 모으는 것이 아니라 각각의 장소에서 데이터마이닝 작업을 통하여 정보를 분석하고 그 분석된 정보만을 중앙장소에 모아서 다시 정보를 분석하는 방법이다. 이 방법은 각 장소에서 분석된 결과만 중앙으로 이동하기 때문에 데이터 이동 비용이 적게 든다. 또 원본 데이터들의 이동이 필요 없기 때문에 보안에 대한 문제점이 없게 된다. 그러나 이를 통한 분석 결과가 모호하거나 부정확할 수도 있다. 이런 단점을 극복하기 위한 다양한 연구가 시도 되고 있는데 이러한 연구들에는 메타학습기법(meta-learning)[1], 지식정보 탐색기술(knowledge probing)[2], 전문가 정보 혼합 방법(mixture of experts)[3], 베이지언 모델 평균법(Bayesian model averaging)[4], 스택 일반화 기법(stacked generalization)[5] 등이 있다. 이와 같은 연구들은 데이터마이닝을 분산된 환경에서 가능하게 하지만 정확한 최종결과를 생성하는 것은 매우 어려운 일이다. 그리고 일반적으로 분산된 위치에 있는 데이터베이스들을 다루기 위해서는 이종 데이터베이스에 관한 문제도 해결해야 하지만 위 연구들은 이종 데이터베이스에 관해 다루고 있지 않다. 그러나 본 논문에서 제시하는 분산형 데이터마이닝 방법론을 이용하여 분석하였을 경우 중앙집중식 데이터마이닝과 동일한 최종결과 생성이 가능 하고 분석시간이 절약된다.

2. 분산형 데이터마이닝 알고리즘 및 시스템

한 곳에 데이터들을 수집할 수 없는 환경에서 분산된 데이터들을 함께 분석하기 위해서는 기존의 데이터 마이닝 알고리즘을 그대로 적용하는 것이 불가능하다는 전제조건하에 본 논문에서는 분산된 환경에서의 데이터 분석이 가능한 데이터 마이닝 알고리즘을 제시하였다. 보편적인 데이터마이닝 기법 중 하나인 의사결정나무 알고리즘을 분산된 환경에 적합하게 고안하였고, 본 알고리즘의 실현성을 증명하기 위해 분산형 데이터마이닝 시스템을 구현하였다.

현실세계의 데이터는 일반적으로 그림 1과 같이 분산되어 있다. 분산된 지역에서 데이터베이스가 각각 존재

* 본 논문은 신기술 연구개발 지원 사업의 지원에 의하여 이루어진 것임(과제번호-10643)

할 때 기존의 의사결정나무 알고리즘을 이용 할 경우 데이터를 한곳으로 이동을 해야만 한다. 이는 의사결정나무 분기에 필요한 필드가 사이트A에는 필드 B, C가 있고 사이트B에는 D, E 가 있어서 한 사이트에 있는 정보만으로는 최종결과를 만들어 낼 수 없기 때문이다. 이와 같은 분산 환경에서의 전제조건은 기존의 의사결정나무 알고리즘을 이용하여 데이터마이닝을 수행하기 위해 분산된 데이터를 한곳으로 모아야만 한다는 것인데, 이동이 불가능한 데이터가 있다면 기존의 의사결정나무 알고리즘을 이용하여 분석을 할 수 없게 된다. 이처럼 분산 환경에서는 기존의 중앙집중식 의사결정나무 알고리즘을 이용할 수 없기 때문에 새로운 알고리즘이 필요하다. 이를 위해 본 논문에서는 분산 환경에 적합한 의사결정나무 알고리즘을 고안하였다. 본 알고리즘의 특징은 사이트A에서 필드B에 대한 분기정보(왼쪽으로 분기될 경우 0, 오른쪽으로 분기될 경우 1)를 사이트B에 전송하는 것인데 이를 이용하여 사이트B에서 특정한 필드 없이 데이터마이닝을 수행할 수 있다. 그리고 본 논문에서 고안한 알고리즘의 가장 큰 특징은 원본 데이터의 이동 없이도 분산된 환경에서 의사결정나무를 완성하여 최종결과를 얻을 수 있다는 것이다.

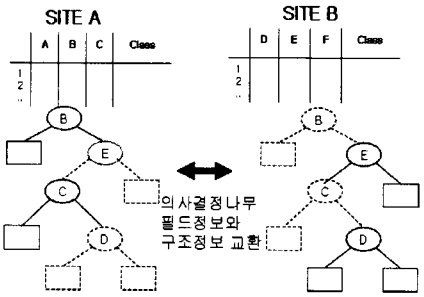

그림 1 분산 환경에서 의사결정나무 모델의 완성

분산형 데이터마이닝 시스템의 구조는 그림 2와 같다. 분산 환경에서 본 시스템은 중앙의 미디에이터와 여러 에이전트 간에 실제 데이터 전송 없이 의사결정나무 모델정보만을 전달하여 최종결과를 생성할 수 있도록 설계되었다. 그리고 정보 분석에 필요한 에이전트를 추가로 참여시킴으로써 시스템 구조의 변화 없이 쉽게 에이전트 수를 증가 시킬 수 있는 장점이 있다. 각 에이전트는 각 사이트에 존재하는 데이터베이스와 데이터마이닝 엔진을 가지고 있지만 각 사이트에 있는 데이터베이스는 분기정보를 모두 가지고 있지 않기 때문에 최종결과를 생성하기에는 완전치 못하다. 따라서 다른 에이전트들과 모델정보를 교환함으로써 최종결과를 생성할 수 있다.

이 시스템은 웹서버, 미디에이터, 에이전트로 이루어져 있는데 웹서버는 사용자가 다른 지역의 데이터베이스를 접근하거나 데이터마이닝 기능을 이용할 수 있도록 웹기반 인터페이스를 지원하고 미디에이터는 에이전트와 에이전트 사이의 통신과 인증을 관리한다. 에이전트들은 각각 다른 사이트에 존재하고 다른 에이전트와 상호 통신을 하여 최종결과를 생성한다.

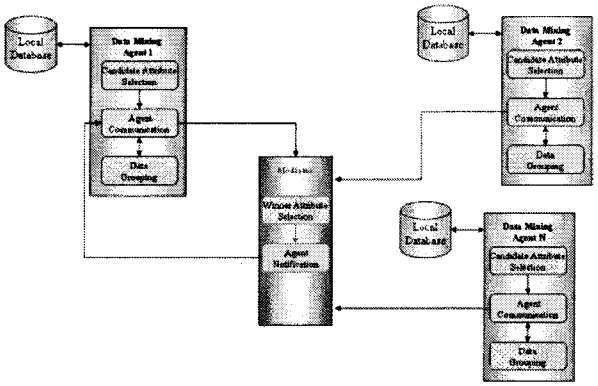

그림 2 분산형 데이터마이닝 시스템 구조

분산 환경에서 에이전트-미디에이터 통신기반 의사결정나무 알고리즘은 다음과 같다.

1.[미디에이터] 각 사이트에 상주해있는 에이전트에게 데이터마이닝 시작을 알린다.

2. [에이전트] 각각의 에이전트는 DB의 필드들 중 의사결정나무의 효율적인 분기를 위한 최적의 필드를 선택함.

3. [에이전트] 선택된 최적의 필드를 미디에이터로 보낸다.

4. [미디에이터] 모든 에이전트로부터 최적의 필드들을 받는 작업이 완료되면 그 필드들을 후보군으로 한 최적 필드 선택작업을 수행.

5. [미디에이터] 선택된 최적의 필드를 가지고 있는 에이전트에게 분기를 시작하도록 알리고 나머지 에이전트는 대기시킴.

6. [에이전트] 선택된 에이전트는 최적의 필드값을 이용하여 분기를 시작함.

7. [에이전트] 분기된 의사결정나무 모델정보(부분결과)를 미디에이터에 전송, 이 때 미디에이터는 모델정보(부분결과)를 다른 모든 에이전트로 전송.

8. [에이전트] 의사결정나무 모델 정보를 이용하여 다른 에이전트들도 모두 똑같이 분기를 수행.

9. [에이전트] 더 이상 분기 될 것이 없을 때, 의사결정나무 규칙을 생성하고 미디에이터에게 종료를 알리고 아직 분기가 끝나지 않았을 때는 2번부터 다시 수행 함.

10. [미디에이터] 모든 에이전트들로 부터 종료신호를 받게 되면 최종적으로 시스템을 종료함.

3. 실험 및 결과분석

본 논문에서 알고리즘의 실용성을 증명하기 위해 90년대 중반 미국 인구통계 데이터(Census bureau Database)[6]를 이용하여 소득 정보에 대한 최종결과를 만들었다. 이 데이터는 매년 미 정부에 의해 실시되며 이 데이터를 이용하여 일반적인 국민의 전체 소득 정보는 각 지역의 경제수준을 비교할 수 있게 한다. 또한 본 논문에서 사용한 인구통계 데이터는 타 논문에서도 사용되는 데이터

로써 이미 데이터의 신뢰성이 확증되었다[7]. 이러한 국가기관에 의해 수집된 방대한 양의 데이터를 데이터마이닝에 이용하게 되면 그 결과는 매우 신뢰도가 높은 정보가 될 수 있다. 따라서 이러한 데이터의 데이터마이닝 결과는 매우 효율적으로 사용될 것이다.

표 1 인구통계 데이터의 애트리뷰트

필드명	설명
AAGE	나이
ACLSWKR	직업
ADTIND	직업 경력(비정규직)
ADTOCC	직업 경력(정규직)
AHGA	최종 학력
AHRSPAY	시급
AHSCOL	현재 학력
AMARITL	결혼 상황
AMJIND	직종(비정규직)
AMJOCC	직종(정규직)
ARACE	인종
AREORGN	히스패닉 계통
ASEX	성별
AUNMEM	노동조합회원
AUNTYPE	실직 사유
AWKSTAT	전시간 또는 단시간 근무 고용 실태
CAPGAIN	자본 소득
CAPLOSS	자본 손실
DIVVAL	주식 배당금
FILESTAT	세금신고 상황
GRINREG	이전 거주 지역
GRINST	이전 거주했던 주
HHDFMX	가족 및 가족구성원 상황
HHDREL	가족구성원 일람
MARSUPWT	-
MIGMTR1	이사 기록(대도시권으로)
MIGMTR3	이사 기록(일반 지역으로)
MIGSUNMIGSAMEMIGMTR4	이사 기록(일반 지역 내)
MIGSAME	현재 거주지에 1년 전에도 거주했는지의 여부
MIGSUN	선벨트 지역에서 거주했는지의 여부
NOEMP	고용된 직원 수
PARENT	18세 이하의 가족 상황
PEFNTVTY	아버지 출생국가
PEMNTVTY	어머니 출생국가
PENATVTY	본인 출생국가
PRCITSHP	미국 시민권자 여부
SEOTR	자영업 또는 프리랜서 여부
VETQVA	-
VETYN	군 가산점
WKSWORK	연간 주간 근무
YEAR	연도
INCOME	소득수준

본 실험에 사용된 인구통계 데이터의 소득 수준을 클래스로 사용하였다. 그러나 소득수준은 문자 값이 아니라 연속된 숫자 값 이므로 클래스로 이용하기 위해서는 기준을 정해야만 한다. 분석결과를 쉽게 확인하기 위하여 기준을 $50000로 정하고 $50000이상은 'higher' 으로, 그 이하를 'lower'로 표기하였다. 그리고 소득수준은 AGI(adjusted gross income)가 아닌 PTOTVAL(total person income) 필드를 이용하였다. 표 2는 본 실험에 사용된 인구통계 데이터의 통계 정보이다.

표 2 인구통계 데이터의 정보

분석 데이터 갯수	129523
중복되는 데이터 개수(분석)	46716
테스트 데이터 갯수	69524
중복되는 데이터 개수(테스트)	20936
'lower' 비율	93.80%
'higher' 비율	6.20%
애트리뷰트(필드) 갯수	40

데이터 프리퍼레이션 작업 이후에는 Learning 파일(.lea)과 Control 파일 (.ctr) 두 개가 생성되어, 예측모델을 생성하기 위해 입력 파일로 사용된다. Learning 파일에는 DB의 속성과 관련된 전반적인 내용이 효율적인 형태로 저장한다(그림 3). Control 파일에는 DB 속성 정보와 클래스 정보, Learning 파일의 내용을 보완할 수 있는 정보와 데이터마이닝에 필요한 옵션 파라미터 등이 포함된다(그림 4).

```
73 3 0 0 12 0 2 6 14 6 4 0 0 1 3 2 0 0 0 4 3 36 29 6 1700.09 0 0 0 1 0
18 3 0 0 0 0 1 4 14 6 1 0 0 1 3 2 0 0 0 4 3 36 7 0 991.95 0 0 0 1 0 0
9 3 0 0 10 0 2 4 14 6 4 0 0 1 3 0 0 0 0 4 3 36 2 2 1758.14 5 6 7 2 2 0
10 3 0 0 10 0 2 4 14 6 4 0 0 1 3 0 0 0 0 4 3 36 2 2 1069.16 5 6 7 2 2 0
8 3 0 0 10 0 2 4 14 6 4 0 0 1 3 0 0 0 0 4 3 36 2 2 2466.24 5 6 7 2 2 0
32 3 0 0 12 0 2 4 14 6 2 0 0 1 3 2 0 0 0 4 3 36 31 6 2021.27 0 0 0 1 0
13 3 0 0 10 0 2 4 14 6 2 0 0 1 3 0 0 0 0 4 3 36 2 2 1520.08 5 6 7 2 2 0
39 3 0 0 0 0 2 2 14 6 4 5 0 1 3 0 0 0 0 2 3 36 36 7 1274.04 5 6 7 2 2 0
16 3 0 0 0 0 0 1 4 14 6 4 6 0 1 3 2 0 0 0 4 3 36 2 2 1555.29 0 0 0 1 0 0
```

그림 3 인구통계 데이터의 LEA파일

```
classes
c: 0 higher
c: 1 lower

attributes
41
a: 0  age con 0.0
a: 1  class_of_worker nom 9 Federal-government
      Local-government Never-worked Not-in-universe Private
      Self-employed-incorporated Self-employed-not-incorporated
      State-government Without-pay
a: 2  detailed_industry_recode con 0.0
a: 3  detailed_occupation_recode con 0.0
a: 4  education nom 17 10th-grade 11th-grade
      12th-grade-no-diploma
      1st-2nd-3rd-or-4th-grade 5th-or-6th-grade
      7th-and-8th-grade
      th-grade Associates-degree-academic-program
      Associates-degree-occup-vocational
      Bachelors-degree(BA-AB-BS) Children
      Doctorate-degree(PhD-EdD)
      High-school-graduate Less-than-1st-grade
      Masters-degree(MA-MS-MEng-MEd-MSW-MBA)
      Prof-school-degree-(MD-DDS-DVM-LLB-JD)
      Some-college-but-no-degree
      ....
      ....
      ....
```

그림 4 인구통계 데이터의 CTR파일

본 논문에서 구현한 시스템을 이용하여 데이터를 분석하면 그 분석결과가 여러가지 형태로 저장된다. 그림 5는 분산형 데이터마이닝을 이용하여 인구통계 데이터를

```
0,0,"lower", "weeks_worked_in_year<45.50 &
 capital_gains<4762.35 & dividends_from_stocks<250.00 &
sex IS Female & member_of_a_labor_union IS NOT Yes &
major_industry_code IS NOT Transportation &
age>=25.30 & major_industry_code IS NOT Agriculture
& tax_filer_stat IS Single &
num_persons_worked_for_employer>=4.95 &
country_of_birth_father IS ? &
marital_stat IS NOT Widowed &
region_of_previous_residence IS Not-in-universe"

1,1,"lower", "weeks_worked_in_year<45.50 &
capital_gains<4762.35 & dividends_from_stocks<250.00 &
sex IS Female & member_of_a_labor_union IS NOT Yes &
major_industry_code IS NOT Transportation &
age>=25.30 & major_industry_code IS NOT Agriculture &
tax_filer_stat IS Single &
num_persons_worked_for_employer>=4.95 &
country_of_birth_father IS ? & marital_stat IS Widowed"

2,0,"higher", "weeks_worked_in_year<45.50 &
 tax_filer_stat IS NOT Nonfiler & capital_gains<4762.35 &
dividends_from_stocks<250.00 & sex IS Female &
member_of_a_labor_union IS NOT Yes & 25.30<=age<33.40 &
major_industry_code IS Agriculture & instance_weight<770.70"

3,1,"lower", "weeks_worked_in_year<45.50 &
tax_filer_stat IS NOT Nonfiler & capital_gains<4762.35 &
dividends_from_stocks<250.00 & sex IS Female &
member_of_a_labor_union IS NOT Yes & 25.30<=age<33.40 &
major_industry_code IS Agriculture & instance_weight>=770.70"
```

그림 5 인구통계 데이터의 의사결정규칙 형태의 분석결과

분석하여 얻어진 의사결정규칙 형태의 분석결과이다. 그림 6은 분산형 데이터마이닝을 이용하여 인구통계 데이터를 분석하여 얻어진 의사결정나무 형태의 분석결과이다.

표 3 인구통계 데이터의 컨퓨징 매트릭스

	ClassA	ClassB
Rule1	1904	2436
Rule2	2655	62529

```
root H = 0.23 49999 3021 46978
    a_39 < 45.50  H = 0.06 30793 325 30468
        a_19 <> 4.00  H = 0.12 12234 318 11916
            a_16 < 4762.35  H = 0.10 12067 252 11815
                a_18 < 250.00  H = 0.07 10775 150 10625
                    a_12 <> 1.00  H = 0.04 6723 45 6678
                        a_13 <> 2.00  H = 0.04 6666 41 6625
                            a_8 <> 21.00  H = 0.04 6623 38 6585
                                a_0 < 25.30  H = 0.01 1142 1 1141  t_1
                                a_0 >= 25.30  H = 0.04 5481  37  5444  t_2
                            a_8 = 21.00  H = 0.25 43  3  40
                                a_24 < 3459.51  H = 0.11 41 1 40  t_16
                                a_24 >= 3459.51  H = 0.00 2  2  0 CLASS 0 RULE 104
                        a_13 = 2.00  H = 0.25 57  4  53
                            a_39 < 39.00  H = 0.00 31 0 31 CLASS 1 RULE 105
                            a_39 >= 39.00  H = 0.43 26  4  22
                                a_0 < 41.50  H = 0.00 13 0 13 CLASS 1 RULE 106
                                a_0 >= 41.50  H = 0.62 13  4  9  t_17
                    a_12 = 1.00  H = 0.12 4052  105  3947
                        a_4 <> 14.00  H = 0.10 3930 85 3845
                            a_0 < 23.28  H = 0.01 769 1 768
                                a_24 < 386.58  H = 0.10 47 1 46  t_18
                                a_24 >= 386.58  H = 0.00 722  0  722 CLASS 1 RULE 115
                            a_0 >= 23.28  H = 0.12 3161  84  3077
                                a_9 <> 2.00  H = 0.11 3065 72 2993  t_19
                                a_9 = 2.00  H = 0.38 96  12  84  t_40
                        a_4 = 14.00  H = 0.45 122  20  102
                            a_8 <> 6.00  H = 0.42 120 18 102
                                a_36 < 1.00  H = 0.47 100 18 82  t_42
                                a_36 >= 1.00  H = 0.00 20  0  20 CLASS 1 RULE 297
                            a_8 = 6.00  H = 0.00 2  2  0 CLASS 0 RULE 298
```

그림 6 인구통계 데이터의 의사결정나무 형태의 분석결과

분산형 데이터마이닝을 이용하여 생성된 최종결과를 테스트하기 위해 9만개의 추가 데이터를 이용하였다. 표 3은 본 논문의 시스템을 이용하여 생성한 최종결과를 테스트한 결과이다. 본 논문의 시스템을 평가하기 위해 인구통계 데이터를 분석한 다른 데이터마이닝 알고리즘의 결과를 참고하였고 본 논문에서 구현한 시스템을 통해 생성한 최종결과를 사용하여 다른 데이터마이닝 알고리즘과의 정확성을 비교하였다. 각각의 데이터마이닝 알고리즘에 따른 에러율은 표 4와 같다.

표 4 각각의 마이닝 알고리즘에 따른 에러율

마이닝 알고리즘	에러율
C4.5	4.8%
C5.0	4.7%
C5.0 rules	4.7%
C5.0 boosting	4.6%
Naive-Bayes	23.2%
자체 알고리즘	7.3%
분산형 데이터마이닝	7.3%

4. 결론

본 논문에서는 분산환경을 위한 의사결정나무 알고리즘을 고안하였고, 알고리즘의 실용성을 증명하기 위해서 분산형 데이터마이닝 시스템을 구축하였다. 시스템의 성능을 평가하기 위해 신뢰성 높은 인구통계 데이터 13만개를 분석하였고 생성된 최종결과의 신뢰도가 다른 마이닝 엔진과 비교를 하였을 때 큰 차이가 없음을 확인하였다. 본 논문에서 제시한 알고리즘을 이용하면 데이터의 위치에 관계없이 모든 데이터를 분석이 가능하고 중앙집중식 최종결과와 같은 결과를 얻을 수 있는 장점이 있다. 데이터마이닝에서 데이터의 양이 분석결과에 가장 큰 영향을 미치고 더 정확한 결과를 얻을 수 있기 때문에 앞으로 분산형 데이터마이닝의 수요는 계속해서 커질 것으로 예상된다.

5. 참고문헌

1. Stolfo, S., Prodromidis, A. L., Tselepis, S. and Lee, W.: JAM: Java Agents for Meta- Learning over Distributed Databases, Proceedings of the International Conference on Knowledge Discovery and Data Mining, pp. 74-81, 1997.
2. Guo, Y. and Sutiwaraphun, J.: Knowledge probing in distributed data mining, In Advances in Distributed and Parallel Knowledge Discovery, 1999.
3. Xu, L. and Jordan, M. I.: Em learning on a generalized finite mixture model for combining multiple classifiers, In Proceedings of World Congress on Neural Networks, 1993.
4. Raftery, A. E., Madigan, D. and Hoeting, J. A.: Bayesian model averaging for linear regression models, Journal of the American Statistical Association, Vol. 92, pp. 179-191, 1996.
5. Wolpert, D.: Stacked generalization, Neural Networks, Vol. 5, pp. 241-259, 1992.
6. http://dataferrett.census.gov/TheDataWeb/index.html
7."Designing a Census Database for Use With GIS", Metropolitan Design Center Technical Paper Series, Number 3, 2005.