Ch.3 데이터 통합(Data Integration)

데이터 통합 (Data Integration)은 여러 데이터 저장소로부터 온 데이터들의 합병을 의미한다. 데이터웨어하우스(data warehouse)나 데이터마이닝과 같은 데이터 분석 작업은 다수의 데이터 원천으로부터 데이터를 하나의 통일된 데이터 저장소로 결합시키는 데이터 통합 작업을 필요로 한다. 데이터 원천은 데이터베이스나 데이터 큐브, 플랫 파일(flat file) 등 다양한 형태로 존재한다. 이제부터 데이터 통합에서 고려해야 할 몇가지 사항에 대하여 살펴보도록 한다.

3.1 개체의 식별

- 개체 식별 문제(Entity Identification Problem)
  - 데이터 통합 시, 동일한 의미의 개체들이 서로 다르게 표현되어 있는 문제 → 어떻게 일치시킬 것인가?
    - ✓ 예) A 데이터베이스에서 customer.customer_id vs. B 데이터베이스에서 cust.cust_number
  - 메터데이터의 역할이 중요
    - ✓ customer_id는 customer 테이블의 PK(Primary Key, 기본키)이고 cust_number는 cust 테이블의 PK이며, 두 속성 다 동일한 데이터 타입과 도메인을 가지고 있다고 한다면, 두 속성은 이름을 다르지만 동일한 속성으로 판단할 수 있음
    - ✓ 일반적으로, 속성의 데이터 타입이나 도메인 뿐만 아니라 기본키 여부, 참조무결성(외래키) 관계, 함수적 종속 관계(functional dependancy) 등을 종합적으로 고려하여 속성의 동일성 여부를 판단해야 함
  - 메터데이터는 데이터변환에도 도움을 줄 수 있음
    - ✓ 예) A 데이터베이스의 성별코드 'M', 'F'vs. B 데이터베이스의 성별코드 '1', '2'
    - ✓ 메터데이터 정보를 이용하여 어느 한 쪽의 데이터변환 필요

3.2 중복

- 중복(Redundancy)
  - 유도속성(derived attribute)
    - ✓ 예) 생년월일과 연령, 월소득과 연간소득, 과목점수와 총점
    - ✓ 월소득 속성값 100만원 vs. 연간소득 속성값 1000만원 ? → 어느 쪽이 틀린 것인가?
  - 정규화되지 않은 테이블
    - ✓ 조회 성능 향상을 위해 일부러 정규화하지 않고 중복 허용 → 일관성 저해의 문제 야기
    - ✓ 예) 구매 테이블 : {구매자번호, 구매일시, **주소**, 전화번호, 구매품목}
    - ✓ 동일 구매자번호에 대해서 다른 주소가 존재할 가능성
  - 중복 문제의 해결은 데이터 정제의 영역으로 어떠한 절대적인 해결책이 있다라기보다는 데이터 정제 시에 가장 데이터 정확성을 높이는 방향으로 정제 룰(cleansing rule)을 정의하여 일괄 적용할 수 밖에 없음

3.3 상관분석

- 상관분석을 통한 중복 탐지
  - 속성 간에 엄격한 함수적 종속 관계가 성립하지는 않지만, 상관분석을 통해서 한 속성이 다른 속성을 얼마나 강하게 암시하는지를 사용 가능한 데이터를 토대로 측정할 수 있음
  - 두 속성 간에 상관도가 높다면 두 속성을 중복으로 보고 그 중 하나의 속성을 제거할 수 있음

3.3.1 수치형 데이터 : 상관계수(correlation coefficient)

- 수치 속성에 대하여 속성 A와 B의 상관계수

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B} \quad \text{식(3-1)}$$

- N : 투플의 개수
- a_i, b_i : 투플 i에서의 속성 A, B의 값
- $\bar{A}$, $\bar{B}$: 속성 A, B의 평균값
- σ_A, σ_E : 속성 A, B의 표준편차

중복 속성 여부를 판단할 때는 해당 분야 도메인 지식(knowledge)을 충분히 고려해서 최종 판단하는 것이 바람직함

상관계수 결과 값의 범위는 –1에서 +1 사이를 만족한다.($-1 \le r_{A,B} \le +1$) 상관계수 $r_{A,B}$의 해석은 다음과 같다

- $r_{A,B} \ge 0$

속성 A, B는 양의 상관관계(positively correlated)를 가진다. 즉, B 값이 증가함에 따라서 A의 값이 증가한다.

- $r_{A,B} \le 0$

속성 A, B는 음의 상관관계(negatively correlated)를 가진다. 즉, B 값이 증가함에 따라서 A의 값은 감소한다.

- $r_{A,B} = 0$

속성 A, B는 독립적이며 둘 사이에 상관관계가 없다.

3.3.1 수치형 데이터 : 상관계수 - 실무예제

다음은 2013년 전국 주요 지점별 유동 인구 현황의 일부이다. 남자 20대 vs. 여자 20대, 남자 10대 vs. 여자 50대의 상관계수를 구하여 비교하고, 중복 속성으로 판단할 수 있을지 검토해 보시오.

⊙ 데이터 파일 : ch3-1(유동인구수).csv

⊙ 원본 투플 수 : 23,221개

조사일자	시간대	X좌표	Y좌표	행정구역명	남자 10대	남자 20대	남자 30대	남자 40대	남자 50대	여자 10대	여자 20대	여자 30대	여자 40대	여자 50대
2010-06-21	12시~13시까지	343099	417482	대전광역시 서구 월평동	2	24	68	50	31	4	37	64	44	26
2010-06-21	19시~20시까지	343099	417482	대전광역시 서구 월평동	19	44	28	33	21	14	56	49	43	18
2010-06-20	12시~13시까지	343099	417482	대전광역시 서구 월평동	13	33	34	61	55	13	32	29	28	12
2010-06-20	19시~20시까지	343099	417482	대전광역시 서구 월평동	23	33	32	547	129	12	39	13	46	4
2010-06-21	12시~13시까지	343121	417343	대전광역시 서구 월평동	0	9	27	21	6	5	24	20	10	6

출처: 공공데이터포털(www.data.go.kr)

☞ 해답은 ch3-1.ipynb 참고

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정

- 범주형(이산형) 데이터인 경우, 속성 A와 B 사이의 상관관계는 피어슨(Pearson)의 카이제곱(χ^2) 검정에 의해 측정

범주형(이산형) 데이터인 경우, 속성 A와 B 사이의 상관관계는 피어슨(Pearson)의 카이제곱(χ^2)검정에 의해 측정될 수 있다. 속성 A가 c개의 범주 값 $a_1, a_2, ..., a_c$를 취하고, 속성 B는 r개의 범주 값 $b_1, b_2, ..., b_r$을 취한다고 가정하자. 그러면, 속성 A와 B에 의해 구성되는 투플은 c개의 열과 r개의 행으로 구성되는 분할표로 표현될 수 있다. (A_i, B_j)를 속성 A가 a_i를 취하고 속성 B가 b_j를 취하는 투플이라고 할 때, χ^2은 다음과 같이 정의된다.

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r} = \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \qquad \text{식(3-2)}$$

- o_{ij} : (A_i, B_j)에 대한 관측도수(observed frequency; 실제로 존재하는 (A_i, B_j) 투플 수)
- e_{ij} : (A_i, B_j)에 대한 기대도수(expected frequency; 확률적으로 기대되는 (A_i, B_j) 투플 수)

e_{ij}는 다음과 같이 계산된다.

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N} \qquad \text{식(3-3)}$$

- N : 데이터 투플 수
- $count(A = a_i)$: 속성 A에 대하여 a_i를 갖는 투플 수
- $count(B = b_j)$: 속성 B에 대하여 b_j를 갖는 투플 수

식 (3-2)에서의 합계는 분할표상의 $r \times c$ 개의 모든 셀에 대하여 계산된다. 따라서 χ^2 값에 가장 크게 기여하는 칸은 실제 관측도수와 기대도수의 차이가 매우 큰 셀이다.

이러한 χ^2 통계량은 속성 A와 B가 독립이라는 가설을 검증한다. 이 검정은 자유도 $(r-1)\times(c-1)$을 갖는 유의수준에 근거하여 검정한다.

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정

- 카이제곱 통계량을 이용한 범주형 속성에 대한 상관분석 사례

어떤 설문조사에서 1,500명의 사람들을 대상으로 각 사람에 대한 성별과 선호 글의 픽션 여부 간의 상관관계를 분석하고자 한다. 성별 속성값은 'male', 'female'이 있고, 픽션 여부 속성값은 'fiction'과 'non-fiction'이 있다. 성별 속성값 분포는 male:female=300:1,200이고 픽션 여부 속성값 분포는 fiction:non-fiction=450:1,050이다. 이들 속성의 조합에 대한 관측도수를 기록한 2×2 분할표는 다음과 같다.

	male	female	Total
fiction	250	200	450
non–fiction	50	1000	1050
Total	300	1200	1500

표에서 각 셀에 대한 기대도수는 식 (3-3)에 의해 구할 수 있다. 예를 들어, 셀 (female, fiction)의 기대도수는

$$e_{12} = \frac{count(femal) \times count(fiction)}{N} = \frac{1200 \times 450}{1500} = 360$$

이와 같은 방식으로 다른 셀에 대한 기대도수를 구하면 다음과 같은 기대도수 분할표를 얻는다.

	male	female	Total
fiction	90	360	450
non–fiction	210	840	1050
Total	300	1200	1500

이제 식 (3-2)에 의해서 χ^2 통계량을 구하면 다음과 같다.

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93$$

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정

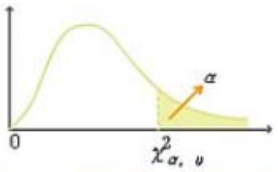
- 카이제곱 통계량을 이용한 범주형 속성에 대한 상관분석 사례

χ^2 통계량을 산출하였으므로 카이제곱 검정 방식에 의해 유의수준 0.05 수준에서의 두 속성 사이에 연관성이 있다라는 가설을 검증해 보자. 본 예제에서의 자유도는 (2−1)×(2−1)=1이 된다. 자유도 1일 때, χ^2 통계량(507.93)에 대한 유의확률(p-value)은 2.2e−16이다(이 수치는 자유도 1일 때의 카이제곱분포로 획득 가능함). 이는 유의수준 0.05보다 훨씬 작은 값으로서 대립가설(두 속성은 연관성이 있다)[7]은 채택된다. 자유도 1일 때 유의수준 0.05로 대립가설을 기각하는 데 필요한 값은 3.842로서 이 값보다 작아야 가설이 기각되는데, χ^2 통계량은 507.93으로 3.842에 비해 매우 크므로 가설은 채택되고, 실제 두 속성 사이에는 강한 연관성이 있다고 결론지을 수 있다.

7) 상관분석 검정의 가설은 크게 귀무가설(두 속성은 연관성이 없다)과 대립가설(두 속성은 연관성이 있다)로 나누어진다.

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정

<카이제곱 분포표>



유의확률(p-value)

자유도(dof)

ν	α=.995	α=.99	α=.975	α=.95	α=.05	α=.025	α=.01	α=.005	ν
1	.3333930	.000157	.000982	.00393	3.841	5.024	6.635	7.879	1
2	.0100	.0201	.0506	.103	5.991	7.378	9.210	10.597	2
3	.0717	.115	.216	.352	7.815	9.348	11.345	12.838	3
4	.207	.297	.484	.711	9.488	11.143	13.277	14.860	4
5	.412	.554	.831	1.145	11.070	12.832	15.086	16.750	5
6	.676	.872	1.237	1.635	13.592	14.449	16.812	18.548	6
7	.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300	12
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819	13
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319	14
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801	15
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267	16
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718	17
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.516	18
19	6.844	7.633	8.907	10.117	30.114	32.852	36.191	38.582	19
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997	20

3.3.2 범주형(이산형) 데이터 : 카이제곱 검정 - 실무예제

다음은 2013년 전라남도 유망중소기업 지정업체 명단의 일부이다. 시군 속성과 지정구분 속성 간의 연관성 여부를 카이제곱 검정 방법에 의해 판단해 보시오(단, 유의 수준은 0.05이다).

⊙ 데이터 파일 : ch3-2(유망중소기업현황).csv

⊙ 원본 투플 수 : 386개

연번	시군	지정구분	기업명	대표자	소 재 지	주생산품	전화번호(061)	비고 (지정번호)
1	목포시	기술유망	브로드컴㈜	이동헌	목포시 석현동 1175(벤처지원 202)	정보통신	284-0017	11월 01일
2	목포시	기술유망	㈜케이에스	김시오	목포시 연산동1236-3	융합철망외	1588-4118	11월 02일
3	목포시	기술유망	삼진물산(주)	김관석	목포시 연산동 1239-1	잡지통조림	270-6113	11월 03일
4	목포시	기술유망	㈜해성	전재두	목포시 연산동1237-3	가드레일외	1588-2811	11월 04일
5	목포시	기술유망	(유)한국메이드	이승룡	목포시 연산동 1238-4	선박블록	278-4411	11월 05일
6	목포시	수출유망	환길산업	박승남	목포시 산정동 1780-1	해조류	272-7147	11월 06일

출처: 공공데이터포털(www.data.go.kr)

☞ **해답은 ch3-2.ipynb 참고**