데이터 매칭 시스템

최창규, 김선기

CONTENTS

01

데이터 매칭

01. 데이터 매칭

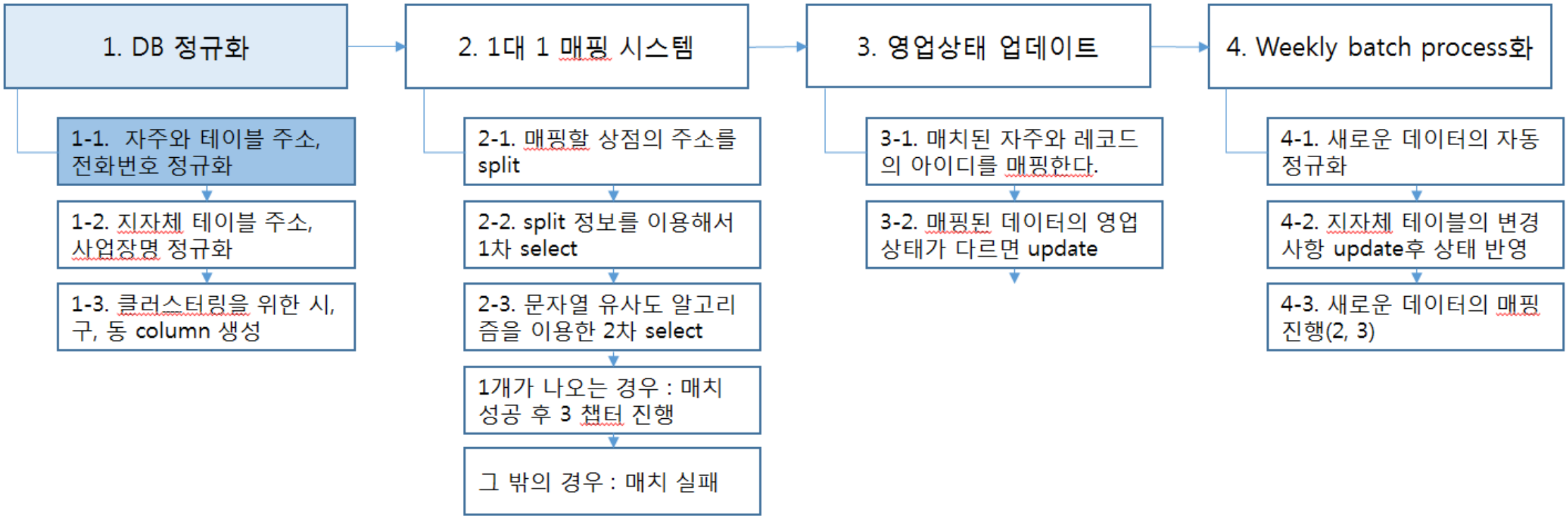
목표 :
자주와 데이터와 지자체 데이터(개폐업 데이터)를 매칭하여 자주와 데이터를 최신화한다.

단계별 방법:
1. DB 정규화
2. 1 대 1 매칭 알고리즘
3. 영업상태 최신화
4. 주간 최신화 작업 자동화

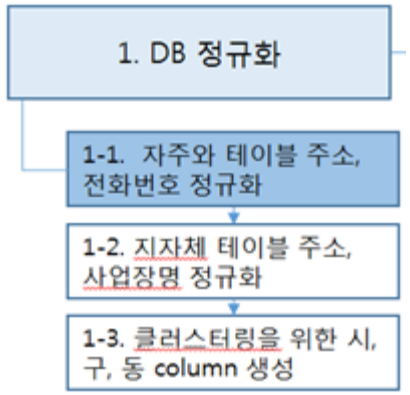
단계별 TASK FLOW : 다음 장

01. 데이터 매칭

매핑시스템 단계별 TASK FLOW-MAP



01. 데이터 매칭



1-1. 자주와 주소/전화번호 정규화
- 대한민국 행정구역 (시구동)양식으로 정제한다.
  # 고도화 단계에서 세부 알고리즘 제작

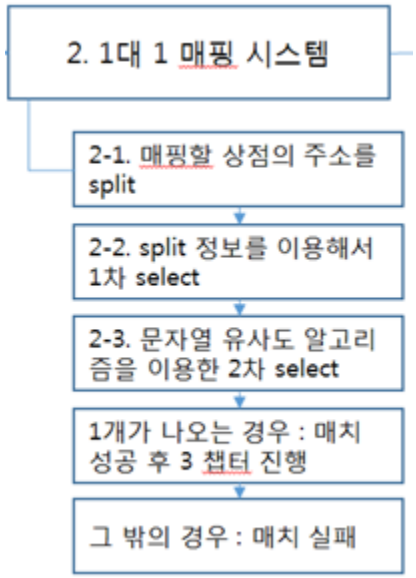
1-2. 지자체 주소/사업장명 정규화 (상동)

1-3. 지자체 주소에서 시/구/동을 추출한다.
- 컬럼 "시" : 광역지방자치단체 (특별시, 광역시, 도, 자치도,자치시)
- 컬럼 "구" : 기초지방자치단체(자치구,군,자치시,특정시)
- 컬럼 "동" : 비자치구역(동, 읍/면,리,일반구, 행정시)

대한민국의 행정 구역

광역지방자치단체	기초지방자치단체	비자치구역		
특별시	자치구		행정동	통
광역시	군		읍·면	행정리
도	자치시		읍·면 행정동	행정리 통
	특정시	일반구		
특별자치도		행정시		
특별자치시				

01. 데이터 매칭



2-1. 자주와 상점주소를 행정구역 단위로 쪼갠다.

2-2. 1-3의 시/구/동 칼럼을 활용해 자주와 시/구/동과 일치하는 지자체 데이터를 가져온다.

2-3. 시/구/동이 일치하는 "상점명/주소"의 유사도를 계산한다.
- 상점명/주소 각각 jaro-distance 알고리즘으로 문자열 유사도 계산
- 두 유사도의 스칼라곱이 0.7이상인 경우, 매칭성공으로 가정한다.

Threshold가 0.7인 이유 : 동일 상점명 또는 동일주소가 각각 다른 상점인 경우가 존재하여 이를 제거할 수치로 임의지정함

2-4. 매칭결과 (다음장)
- 100개 Random Sampling 10번 시도한 결과, 약 30%의 매칭성공률
- Threshold 낮추면 False Positive ratio 높아짐
- 자주와/지자체 데이터의 정제 고도화 선행에 초점

01. 데이터 매칭

CU신대지구점 전라남도 순천시 해룡면 신대리 2004 0617278363
CU신대지구점 전라남도 순천시 해룡면 신대리 2004번지 061-727-8363 0.96

C누크당구클럽 강원도 홍천군 홍천읍 신장대리 77-3 0334320002
C누크당구클럽 강원도 홍천군 홍천읍 신장대리 77-3번지 432-0002 0.96

건강민물매운탕 경기도 평택시 포승읍 방림리 54-11 0316810898
건강민물매운탕 경기도 평택시 포승읍 방림리 54-4번지 031 6810898 0.9

곰내노래연습장 경상남도 창원시 진해구 남문동 868-3 0555451203
곰내 경상남도 창원시 진해구 남문동 868번지 545-1203 0.7371000000000001

금학칼국수 강원도 평창군 봉평면 창동리 359-11 0333351777
금학칼국수 강원도 평창군 봉평면 창동리 359-11번지 3351777 0.97

나이아가라모텔 경기도 포천시 이동면 도평리 60-1 0315355932
나이아가라 경기도 포천시 이동면 도평리 60-1번지 5350164 0.9023999999999999

남부다실 경상북도 문경시 모전동 65-23 0545535258
남부다실 경상북도 문경시 모전동 65-23번지 054 5535258 0.94

대경장모텔 울산광역시 울주군 삼남면 교동리 1499-140 0522633456
대경장여관 울산광역시 울주군 삼남면 교동리 1499-140번지 052 2632355 0.7857000000000001
...

데이터 매칭 알고리즘을 이용해서 매칭된 페어 중 일부(위 : 자주와, 아래 : 지자체)

02

주소 정제 고도화

02. 주소 정제 고도화

자주와 **지번주소기준 오류현황 파악**

총 4,789,219개 중 3,788,215개 오류 (79.1%)

오류 유형	개수(비율)
광역자치구역 주소 오류 - 서울 서초구 서초동	3,647,028 (96%)
상세주소 누락 오류 - AA시 BB구 CC동, AA시 BB구 (세종시 제외) - 일부 상세주소가 있어도 완전한 주소 아니므로 오류판단	130,890 (3%)
도로명주소가 포함된 오류 - AA시 BB구 CC동 DD대로, AA시 BB구 DD대로 - 도로명이 섞였거나, 도로명으로만 구성된 주소	22,655 (0.6%)
주소정보가 없는 오류 - NULL , 123-24번지, 산25	827(0.02%)
주소 오탈자 - 고양시 일산동구 (광역자치구역 주소 부족) - 경기도 고량시 (행정구역 오타) - 대구남구북동 (띄어쓰기X)	57(0.00%)

02. 주소 정제 고도화

자주와 오류유형별 고도화 방법

오류 유형	개선방안
광역자치구역 주소 오류 - 서울 서초구 서초동	정답주소 : 오류주소(1:N)의 관계사전을 활용해 정규화
상세주소 누락 오류 - AA시 BB구 CC동, AA시 BB구 (세종시 제외) - 일부 상세주소가 있어도 완전한 주소 아니므로 오류판단	'mygeoposition.com'사이트에서 좌표 검색결과 주소(크롤링)로 변환
주소정보가 없는 오류 - NULL , 123-24번지, 산25	
도로명주소가 포함된 오류 - AA시 BB구 CC동 DD대로, AA시 BB구 DD대로 - 도로명이 섞였거나, 도로명으로만 구성된 주소	네이버/다음 지도 검색결과 주소를 크롤링 (API활용)
주소 오탈자 - 고양시 일산동구 (광역자치구역 주소 부족) - 경기도 고량시 (행정구역 오타) - 대구남구북동 (띄어쓰기X)	대쉬보드를 통해 수정

02. 주소 정제 고도화

오류유형별 고도화 방법 : 정답주소-오류주소(1:N)의 관계사전을 활용해 정규화

```
addr_list['서울특별시']=('서울특별시', '서울시', '서울')
addr_list['경기도']=('경기도', '경기')
addr_list['강원도']=('강원도', '강원')
addr_list['충청북도']=('충청북도', '충북')
addr_list['충청남도']=('충청남도', '충남')
addr_list['경상북도']=('경상북도', '경북')
addr_list['경상남도']=('경상남도', '경남')
addr_list['전라북도']=('전라북도', '전북')
addr_list['전라남도']=('전라남도', '전남')
addr_list['부산광역시']=('부산', '부산광역시', '부산시')
addr_list['대구광역시']=('대구', '대구광역시', '대구시')
addr_list['인천광역시']=('인천', '인천광역시', '인천시')
addr_list['광주광역시']=('광주', '광주광역시', '광주시')
addr_list['대전광역시']=('대전', '대전광역시', '대전시')
addr_list['울산광역시']=('울산', '울산광역시', '울산시')
addr_list['세종특별자치시']=('세종', '세종시', '세종특별시', '세종특별자치시')
addr_list['제주특별자치도']=('제주', '제주도', '제주특별도', '제주특별자치도')


sub_addr_list['충청북도']=('제천시', '충주시', '단양군', '음성군', '진천군', '증평군', '괴산군', '청주시', '보은군', '옥천군', '영동군')
sub_addr_list['충청남도']=('논산시', '계룡시', '서산시', '태안군', '공주시', '부여군', '천안시', '홍성군', '예산군', '아산시', '서천시',
                           '당진시', '보령시', '청양군', '금산군')
sub_addr_list['경상북도']=('포항시', '경주시', '김천시', '안동시', '구미시', '영주시', '영천시', '상주시', '문경시', '경산시', '군위군',
                           '의성군', '청송군', '영양군', '영덕군', '청도군', '고령군', '성주군', '칠곡군', '예천군', '봉화군', '울진군', '울릉군')
sub_addr_list['경상남도']=('창원시', '김해시', '진주시', '양산시', '거제시', '통영시', '사천시', '밀양시', '함안군', '거창군', '창녕군', '고성군',
                           '하동군', '합천군', '남해군', '함양군', '산청군', '의령군')
sub_addr_list['전라북도']=('전주시', '익산시', '군산시', '정읍시', '김제시', '남원시', '완주군', '고창군', '부안군', '임실군', '순창군', '진안군',
                           '무주군', '장수군')
sub_addr_list['전라남도']=('목포시', '여수시', '순천시', '나주시', '광양시', '담양군', '곡성군', '구례군', '고흥군', '보성군', '화순군', '장흥군',
                           '강진군', '해남군', '영암군', '무안군', '함평군', '영광군', '장성군', '완도군', '진도군', '신안군')
```

02. 주소 정제 고도화

오류유형별 고도화 방법 : 크롤링

Case 1 : 주소정보가 없거나 상세주소 누락된 경우

>> 좌표 정보를 이용해서 주소로 변환 (http://mygeoposition.com/)

Case 2 : 도로명 주소로 되어있거나 도로명 주소와 지번주소가 섞인 경우

>> 네이버, 다음 지도에서 지번주소로 변환
- 단순 크롤링시 제한사항
  : 클래스명 불규칙 변동, 서버접근LOCK
- 네이버 API 사용

네이버지도 : 커피빈 교대점

02. 주소 정제 고도화

오류유형별 고도화 방법 : 대쉬보드 상에서 수동 수정 반영

ID	CO_NAME	ADDR	LATITUDE	LONGITUDE
286141	뚜레쥬르발산2동점	서울특별시 강서구 내발산동 703	37.5525304438664	126.837231442889
286212	뚜레쥬르소하동양점	광명시 소하동 48-56 (지번)	37.4541606557629	126.88857530123
286360	뚜레쥬르익산원광대점	익산시 신동 763-2	35.9637100976929	126.957683046457
379128	뚜레쥬르 안성죽산점	안성시 죽산면 죽산리 560-5	37.0734886965614	127.418131213022
515972	세븐일레븐 대구논공점	대구시달성군 논공읍 용로로1길36	35.729060191513	128.449399086278
516018	세븐일레븐 대구미래점	대구남구 두류공원로 22(대명동)1층	35.8406515477232	128.574789438104
516074	세븐일레븐 대구서부시외터미널점	대구남구 월배로 496(대명동)	35.8366615537929	128.557065578392
516152	세븐일레븐 대구진천역점	대구달서구 월배로 64(진천동)1층 2:	35.8134276452243	128.521774469898
516327	세븐일레븐 대전 용두점	대전광여시 중구 계룡로 815(용두동	36.3274244413886	127.406928488557
516466	세븐일레븐 도화마이다스점	. 인천광역시 남구 제일로31	37.4592774555805	126.673900346283
516597	세븐일레븐 마산양덕중앙점	경남남도 창원시 마산회원구 양덕남	35.2223501704382	128.587505505736

주소 오탈자 검색 대쉬 보드

02. 주소 정제 고도화

detect한 모든 데이터는 정제되었다.

- 광역자치주소 .. 품질 미확인 상태 : 해결과제