Crowdsourcing Data Understanding: A Case Study using Open Government Data

Yukino Baba, Hisashi Kashima (Kyoto Univ.)

HIGHLIGHT

We investigate the feasibility of crowdsourcing for data understanding

RESULT1: Crowds can have sufficient skills to provide reasonable findings and 79% of the findings are correct

RESULT2: Crowds generate diverse findings and 87% of the findings are not overlapped with others.

BACKGROUND

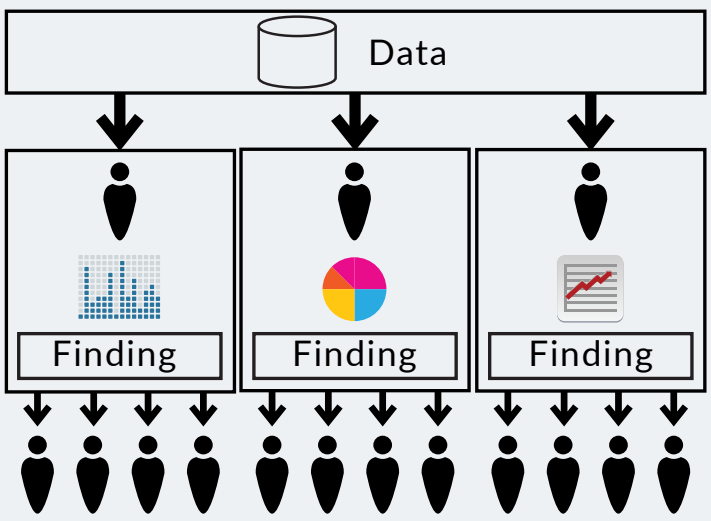
Crowdsourcing is a promising solution to compensate for the lack of data analysts

Business understanding → **Data understanding** → **Data preparation** → **Modeling** → **Evaluation/ Deployment**

Crowdsourcing has been already used for Data preparation and Modeling

PROPOSED WORKFLOW

Workers are asked to provide both finding and supporting chart



STEP1: Data exploration
Crowds provide findings and supporting chart

STEP2: Review
Crowds review findings

CASE STUDY

Open government data of Japan was used

E.g., Statistics about road traffic

Trends in fatalities by month

month / year	January	February	March	April	May	June	The first half of the year	July	August	September	October	November	December	the latter half of the year	The year total
1970	1,237	1,140	1,379	1,271	1,419	1,289	7,735	1,480	1,545	1,467	1,476	1,515	1,547	9,030	16,765
2000	728	667	780	697	695	697	4,264	747	806	686	835	867	868	4,809	9,073
2001	619	637	764	666	663	662	4,011	743	745	726	824	834	874	4,746	8,757
2002	648	633	736	691	642	623	3,973	656	698	670	762	801	836	4,423	8,396
2003	597	560	625	573	609	573	3,537	586	711	644	740	748	802	4,231	7,768
2004	561	517	624	611	587	563	3,463	640	627	587	649	692	767	3,962	7,425
2005	563	472	573	531	499	511	3,149	582	614	637	616	655	674	3,778	6,927
2006	535	426	555	489	474	469	2,948	527	569	509	548	650	652	3,455	6,403
2007	495	451	452	423	430	427	2,678	473	527	475	549	508	572	3,104	5,782
2008	403	361	388	402	387	371	2,312	449	475	398	502	491	570	2,885	5,197
2009	384	364	387	357	404	352	2,248	380	438	405	467	489	541	2,720	4,968
2010	393	352	366	353	380	354	2,198	407	434	412	469	425	577	2,724	4,922
2011	331	360	381	370	346	343	2,131	363	408	378	471	429	483	2,532	4,663
2012	324	322	341	337	309	301	1,934	344	392	368	438	433	502	2,477	4,411
2013	345	336	333	345	332	313	2,004	331	373	366	378	431	490	2,369	4,373
2014	355	307	311	313	322	317	1,925	325	301	345	400	377	440	2,188	4,113
2015	346	308	317	320	315		1,606								1,606
change	-9	1	6	7	-7		-2								-2
percentage change	-2.5	0.3	1.9	2.2	-2.2		-0.1								-0.1
compared with 2014															
Fatalities per day	11.2	11.0	10.2	10.7	10.2		10.6								10.6

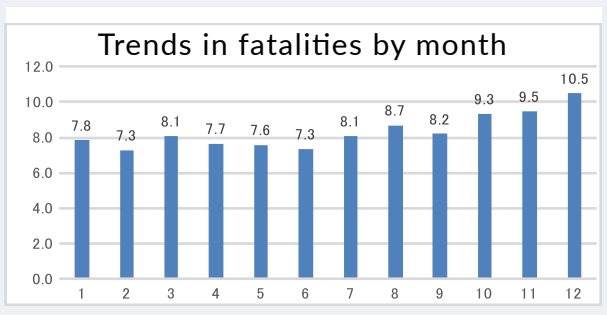
A worker is asked to provide three finding-chart pairs

RESULTS

Crowdsourcing is efficient in both data exploration and review tasks

Example of finding: "the average number of deaths in traffic accidents over the last 15 years was the highest in December"

Example of chart:



Confusion matrix of worker review

79% of findings were correct

		Crowd review: Correct	Incorrect	No-graph	Total
Expert review	Correct	90	0	0	90
	Incorrect	17	5	0	22
	No-graph	0	0	2	2
	Total	107	5	2	114

Precision was 84% Recall was 100%

There were 97 unique findings and 87% of them were not overlapped with others