Published in final edited form as:

Proc IEEE Int Symp Comput Based Med Syst. 2014 May ; 2014: 347–350. doi:10.1109/CBMS.2014.79.

Computer-Based Coding of Occupation Codes for Epidemiological Analyses

Daniel E. Russ,

Division of Computational Bioscience, Center for Information Technology, NIH, Bethesda, MD, 20892 USA

Kwan-Yuet Ho,

Division of Computational Bioscience, Center for Information Technology, NIH, Bethesda, MD, 20892 USA

Calvin A. Johnson, and

Division of Computational Bioscience, Center for Information Technology, NIH, Bethesda, MD, 20892 USA

Melissa C. Friesen

Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD, 20892 USA

Daniel E. Russ: druss@mail.nih.gov; Calvin A. Johnson: johnson@mail.nih.gov; Melissa C. Friesen: friesenmc@mail.nih.gov

Abstract

Mapping job titles to standardized occupation classification (SOC) codes is an important step in evaluating changes in health risks over time as measured in inspection databases. However, manual SOC coding is cost prohibitive for very large studies. Computer based SOC coding systems can improve the efficiency of incorporating occupational risk factors into large-scale epidemiological studies. We present a novel method of mapping verbatim job titles to SOC codes using a large table of prior knowledge available in the public domain that included detailed description of the tasks and activities and their synonyms relevant to each SOC code. Job titles are compared to our knowledge base to find the closest matching SOC code. A soft Jaccard index is used to measure the similarity between a previously unseen job title and the knowledge base. Additional information such as standardized industrial codes can be incorporated to improve the SOC code determination by providing additional context to break ties in matches.

Keywords

Automated Coding; Occupational Coding

I. Introduction

Characterizing occupation is a crucial first step in epidemiological analyses of occupational risk factors. It is also crucial in the evaluation of exposure patterns by time, industry, and occupation based on data found in exposure databases, such as inspection databases. In both contexts, job title is generally provided as free-form text responses to open-ended questions

(e.g. “What is your job title?”). Occupation information is mapped from text to Standard Occupational Classification (SOC) codes [1] to group similar jobs together to facilitate exposure modeling and epidemiologic analyses. The SOC codes are hierarchical. The first 2 digits are the major group; the first 3 digits are the minor group of occupations; and all 6 digits represent a detailed occupation.

Manual coding of job titles is impractical for very large datasets required in occupational exposure studies. Automated mapping of job titles to SOC codes provides an important first step in understanding the exposure patterns based on occupation.

Automatic mapping of job titles is a challenging problem. Job titles are short text phrases, confounding natural language processing. Spelling errors and non-standard abbreviations in the text can cause direct matching of job titles to fail. In addition, often a job title is not specific enough to differentiate between SOC codes (e.g. consider “*technician*”).

Other systems for computer-based occupational coding have been developed. These systems are used by US state and federal agencies interested in unemployment insurance claims. One system, *OccuCoder* [2], is available to state unemployment agencies for coding resumes, jobs, and unemployment claims. A web-based trial version capable of coding one job title is available.

In this paper, we focus on the methodology used to identify SOC codes from free text job titles. In Section 2, we first provide a brief overall view of the solution, followed by an explanation of the data and techniques used. Section 3 presents the results of a small preliminary comparison between human coders and our computer-based system. The results from *OccuCoder* are also presented for comparison.. In Section 4, we discuss the results and identify additional research that may improve results.

II. Methodology

A. Overview of System

At the core of our computer-based coding system is a k -Nearest Neighbor (kNN) classifier. A large knowledge base of job titles with associated SOC codes was created with publically available data. Query job titles are compared to the knowledge base to find the SOC codes of the most similar job titles in the knowledge base. In our study, we chose ($k=1$) or the SOC code of the most similar job title in the knowledge base. The similarity between known job titles is measured using a soft Jaccard index to handle misspellings and unconventional abbreviations. Additionally, we attempted to improve coding performance by using additional information present in the dataset.

B. Knowledge base development

Job title information was taken from various sources. The O*NET production database [3] contains example job titles for each SOC code, and crosswalks [4] mapping additional job titles to SOC codes. Additional job titles came from the Direct Match Title File [1] provided by the Bureau of Labor Statistics, and 2010 U.S. Census Occupational Index [5]. Combining all of the sources, approximately 93,000 job titles are in our knowledge base.

C. IMIS data set

The Integrated Management Information System (IMIS) is a very large exposure database of inspection measurements of U.S. work sites maintained by the U.S. Occupational Safety and Health Administration [6]. The IMIS dataset is publically available via a Freedom of Information Act request for a handling fee. Our dataset is a subset of the IMIS data set, containing lead and cadmium exposure data from over 80,000 personal measurements, with over 17,100 unique free-text job titles. In addition to job titles, Standard Industrial Classification [7, 8] (SIC) codes are available for each measurement, which identify the industry of the employer.

D. Job Title String Preprocessing

The job titles in the IMIS dataset are hand entered free text containing misspelling and unnecessary descriptors to the job title (e.g. welder #1). Non-alphabetic characters are replaced with spaces. Consecutive spaces are condensed into a single space.

E. Job Title Similarity

Similarity between two job titles was measured using a soft Jaccard index. A hard Jaccard index is defined as the intersection of two sets divided by the union. In our case, each job title was tokenized. The number of words common to both job titles is the intersection. The union is the total number of unique words that occur when the job titles are combined. The hard Jaccard fails because slight variants of a word, including slightly misspelled words, are considered unrelated. The soft Jaccard index allows similarly spelled words to partially intersect by using a membership function.

We developed a membership function based on the Damerau-Levenshtein (DL) distance and the longest common prefix. The soft Jaccard recognizes slightly misspelled words. Highly misspelled words are not affected, to avoid unintentionally changing the concept not just the spelling. The longest common prefix (LCP) handles non-standard abbreviations (e.g. tech for technician). A comparison between the soft and hard Jaccard is shown in Fig. 1 for the two phrases “junior system analyst” and “systems analyst”. The hard Jaccard index, shown in Fig. 1A, treats “system” and “systems” as completely separate words. Only one of four words overlap and the Jaccard index is 0.25. The soft Jaccard, shown in Fig. 1B, allows “system” to partially overlap “systems”. A full overlap would result in two out of three words in common resulting a Jaccard index of 0.67. The resulting soft Jaccard index is 0.62, very close to a complete overlap. Our membership function is defined as

$$m = \max \left\{ \frac{\max(l_1, l_2) - DL}{\max(l_1, l_2)}, \frac{LCP}{\max(l_1, l_2)} \right\} \quad (1)$$

where m is the membership, l_1 and l_2 are the lengths of words one and two respectively, LCP is the length of the longest common prefix, and DL is the Damerau-Levenshtien distance. A word in one job title is paired with only one word from the other job title. The words with the largest membership are paired first. In the example from Fig. 1, after the word “system” is paired with “systems” neither word can be used again. In our

implementation, the *DL* distance between two words is calculated using LingPipe [9], a royalty-free text processing toolkit.

F. Adding Additional Features

The SIC code information in the IMIS data was used as an additional feature during classification. Industrial information available from the National Industry-Occupation Employment Matrix provides the prevalence of SOC codes for each industry. Because the industry information was coded with North American Industry Classification System (NAICS), a SIC to NAICS crosswalk was used to map the SIC code to potential NAICS codes. The SIC prevalence is the conditional probability of a SOC code given a particular SIC code, which can help reduce the score of incorrect matches. An overview of the calculation to calculate the most probable SOC is shown in Fig. 2. An average rank was calculated based on the rank order from the soft Jaccard score and the rank order based on most probable SOC and was used to identify the best SOC code based on SIC and job title.

G. Manual Coding of Job Titles

In our preliminary study, two independent coders familiar with occupational coding systems manually coded 100 job titles. Each job title had a SIC code to provide industry information. Using a web-based application that provided a list of all the job titles/SIC codes, and the SOC hierarchy, the coders were allowed to choose up to 3 choices for the SOC code, and they were given a special code for unknown/not able to code (e.g. Job title of *n/a*). The application recorded the coder's selections.

Inter-rater agreement was measured using Cohen's kappa [10]. Even though the coders were allowed to select multiple SOC codes for a given job title, for the calculation of Cohen's kappa only one SOC code per job title is used. When the coders agree, only the matching SOC codes are used to calculate inter-rater agreement, regardless of whether it is the first, second, or third choice. When the coders disagree, only the first choices are used.

Job titles were split into six categories based on the level of coder agreement to reflect the level of difficulty in classifying the job titles. In the first and second categories, the coders agree on the SOC at a 6-digit and 3-digit level respectively with their first choice. In the third and fourth categories, the coders agree at the 3- or 6-digit level with a second choice. These categories include cases where both of the coders agree with their second choices, or if a first choice matches the other's second choice. The fifth category is where the coders fail to agree at the 3- or 6-digit level.

III. Results

The results of the preliminary manual coding are shown in Table I, along with illustrative examples. The coders agreed on 21 (24%) job titles at the 6-digit level (including matches to first or second choices). Cohen's kappa at the 6-digit level was 0.35. At the 3-digit level, coders agree on 66 (74%) job titles. Matches at the 6-digit level are included in the 3-digit level. The Cohen's kappa at the 3-digit level is 0.67.

Comparisons between our system and the coders are shown in Table II. The system results are provided with and without using the SIC information. Also shown in Table II are the results of *OccuCoder*. The coders agreed with their first choices at the 6-digit SOC level 15 times. For those 15 job titles, the system agreed with the coders 10 times at the 6-digit level and 12 times at a 3-digit level. When the coders disagreed, a system match to either coder was considered a match. The 11 job titles that the coders were unable to assign codes were not included in the analysis.

We looked at the 32 job titles where the system disagreed with the coders at the 3-digit level. Out of these job titles, 19 (59%) were cases where the soft Jaccard allowed spelling changes from one word to a completely different concept (e.g. solder to loader).

Adding the SIC code by average rank decreased performance of the coding in all cases except when the coders agreed in the 2nd choice at the 3-digit SOC level, which had no improvement.

The *OccuCoder* results are shown without using SIC information, because the web interface did not use the SIC.

IV. Discussion

We developed a system that automatically assigns SOC code given a job title. At first, the preliminary results of the small-scale study (64% agreement rate at the 3-digit SOC level) may appear to be disappointing; however, the human raters agreed on only 66 (74%) of the 89 job titles that were coded. Coders were allowed to select SOC codes at any level in the SOC hierarchy, whereas the system could only pick at the 6-digit level. Selecting a SOC code higher in the hierarchy guarantees the system will not match.

In this preliminary study, our system performed as well as *OccuCoder*. At almost every coder agreement level, our system matched more job titles with the coders than *OccuCoder*. Even though the difference was not large enough given the size of the test to claim better performance (t -test $p > 0.05$), our initial results are very promising.

Our preliminary analyses provided several suggestions of where the automatic coding could be improved for future use. Soft matching correctly spelled words led to many of the systems coding errors. The system can improve performance by detecting correctly spelled words before soft matching. Nineteen incorrectly coded job titles had soft match with correctly spelled words. If half of the concept changes were corrected, the performance of the system would on par with the performance of the human coders. Other sources of error still need to be studied.

Our initial attempt to integrate the SIC information did not improve performance. Average rank was too naïve and it led to decreased performance. However, human coders found industry information very useful when choosing between related SOC codes. Logistic regression modeling may provide better results than average rank. Additional features may also be required by the regression, for example, the number of SOC codes with exactly matching job titles.

In addition to improving the soft Jaccard measure and SIC integration, a larger round of validation is underway which will contain a consensus assignment for job titles where the coders disagreed. The larger test dataset will provide a high quality gold standard for comparison that will allow us to determine the properties that predict a good match, so that we may prioritize human coders to review only those predicted to be a poor match.

Acknowledgments

This research was supported by the Intramural Research Programs of the Center for Information Technology and the National Cancer Institute.

References

1. U.S. Bureau of Labor Statistics. Standard Occupational Classification System (SOC) System. 2014 Jan 15. [Online]. Available: <http://www.bls.gov/soc>
2. Information Technology Service Center. OccuCoder Product Information. 2014 Mar 24. [Online]. Available: http://www.itsc.org/Pages/pub_aocode.aspx
3. O*NET Resource Center. O*NET 18.0 Production Database. 2014 Jan 15. [Online]. Available: <http://www.onetcenter.org/database.html>
4. O*NET Resource Center. Supplemental Database files. 2014 Jan 15. [Online]. Available: <http://www.onetcenter.org/supplemental.html>
5. U.S. Census Bureau. Industry and Occupation – Indexes – People and Households – U.S. Census Bureau. 2014 Jan 15. [Online]. Available: <http://www.census.gov/people/io/methodology/indexes.html>
6. Lavoue J, Friesen MC, Burstyn I. Workplace measurements by the U.S. Occupational Safety and Health Administration since 1979: Descriptive analysis and potential uses for exposure assessment. *Ann Occup Hyg.* Jun.2013 57:681–3. [PubMed: 23946941]
7. United States. Office of Management and Budget. Standard industrial classification manual. Washington, D.C: 1987.
8. Occupational Safety and Health Administration, US Department of Labor. SIC Division Structure. 2014 Jan 15. [Online]. Available: http://www.osha.gov/pls/imis/sic_manual.html
9. Alias-I. LingPipe 4.1.0. 2008. [Online]. Available: <http://alias-i.com/lingpipe>
10. Cohen J. A Coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* 1960; 20:37–46.

Fig. 1.

Comparison between hard Jaccard index and soft Jaccard index. The hard Jaccard (A) treats the words “system” and “systems” as two separate tokens. Only 1 of the 4 tokens intersect yielding a Jaccard index of 0.25. The soft Jaccard (B) allows “system” and “systems” to become the same token, with a membership function of 0.86 in set X and 1 in set Y . The soft Jaccard index is 0.62, which is very close to 0.67, the hard Jaccard index had set Y been {“system”, “analyst”}.

Fig. 2.

Identification of the most probable SOC code given a SIC. A crosswalk is used to get the NAICS codes for a given SIC. The industry-occupation matrix can then be used to find the most probable SOC codes.

TABLE I

Example Job titles for the different Agreement Levels

SOC Agreement Level	Example Job Title	Example SIC	Coder 1		Coder 2		Counts at Agreement level
			1 st Choice	2 nd Choice	1 st Choice	2 nd Choice	
6-digit	mixer	3269	47-2061		47-2061	47-2060	15
3-digit	small parts operator	3692	51-2022	51-2090	51-2099		38
6-digit, 2 nd choice	boston pot oper	3691	51-4051		47-2060	51-4051	6
3-digit, 2 nd choice	grinding-rice lake	3321	51-4033	51-9021	51-9020		7
No agreement	Pouring floor labore	3366	51-9198		47-2051	47-4000	23
Unable to code	n/a	7692	99-0000		99-0000		11

TABLE II

Agreement between Human Coders, Our System, and *OccuCoder*

Coders SOC agreement	Count	Job title Only				Job title and SIC				OccuCoder			
		6-Digit Match	3-Digit Match	3-Digit Match	6-Digit Match	6-Digit Match	3-Digit Match	3-Digit Match	6-Digit Match	6-Digit Match	3-Digit Match	3-Digit Match	
6-Digit	15	10	67%	12	80%	6	40%	7	47%	8	53%	10	67%
3-Digit	38	21	55%	23	61%	17	45%	19	50%	12	32%	20	53%
6-Digit 2 nd Choice	6	5	83%	5	83%	3	50%	3	50%	2	33%	3	50%
3-Digit 2 nd Choice	7	4	57%	6	86%	2	29%	6	86%	4	57%	6	86%
No agreement	23	5	21%	11	48%	4	17%	11	48%	6	26%	12	52%
Total	89	45	51%	57	64%	32	36%	48	54%	32	36%	51	57%
Unable to Code	11												