Thai Celebrity Information Extraction Based on Association Rule Measures

Chinorot Wangtragulsang, Nattakarn Phaphoom, Phannachet Na Lamphun, Pisit Charnkietkong, and Jian Qu

Faculty of Engineering and Technology
Panyapiwat Institute of Management, Nonthaburi, Thailand
E-mail: chinorot_w@hotmail.com, nattakarnpha@pim.ac.th, phannachetnal@pim.ac.th, pisitcha@pim.ac.th, jianqu@pim.ac.th

Received: April 22, 2019 / Revised: August 21, 2019/ Accepted: August 29, 2019

***Abstract*— This study aims to develop a system to automatically extract and select celebrity information from websites, as traditionally celebrity information is gathered and selected by hand, which is rather time-consuming and often unable to stay updated due to large number of celebrities in Thailand, and potential ambiguity and conflict between information sources on the Internet. This study proposes a novel method that uses pattern matching and association rules to extract date of birth, height and weight of celebrities from websites. In addition, a weight estimation system based on height and BMI is developed. It is found that our system is able to obtain more celebrity information than many Thai websites such as MThai.com. Also, the weight estimation system is able to estimate celebrities' weights based on height and BMI index.**

***Index Terms*—Information extraction, personal information extraction, unstructured data, weight estimation**

I. Introduction

Personal information of celebrities, such as actors and singers, has been normally collected by hand or in some examples through crowd-sourcing (like Wikipedia), and then stored in database for further use. One notable example is the Internet Movie Database (IMDb), which allows registered members (including the celebrities themselves, their agents, or the production crew) to put in additional data as needed. Due to its long history, ownership by Amazon, and huge number of contributors, the IMDb is able to keep the large amount of information up-to-date.

In non-crowd-sourced websites however, usually the web administrator or staff must perform research, validation and import solely by hand, which although highly accurate, is fairly time-consuming and poorly-scalable against large amount of data. In this case, the computer's role in such data extraction is limited to being a vessel for the human-processed data. This can be partly attributed to lack of punctuation mark to separate words which makes information extraction by machine is a challenging task.

Although there are Thai websites that collect data on Thai celebrities such as Thaiza[1], MThai[2], Siamdara[3] and the Thai version of Wikipedia, such efforts are not centralized and as a result, information on celebrities were fragmented (present on one website but not others). In addition, due to large amount of Thai celebrities, smaller Thai websites have limited data management abilities. One notable example is the Siamdara website where few, if any, Thai celebrity has information post-2015, and MThai does not have as many celebrities as Siamdara.

Due to difficulties in updating large amount of Thai-language celebrity information, we propose a novel automated method to solve above problems. Our method uses novel rules to extract possible pieces of information (date of birth, height and weight) or "personal information candidates" from websites, and then association rule measures and statistical rules are used to select the most likely personal information candidate. In addition, weight information is relatively rare compared to information such as date of birth, especially for female actresses, possibly due to privacy reasons. Furthermore, in an effort to solve the missing weight information problem, we propose a modified BMI method to estimate weight for the celebrities.

The remaining parts of this paper are similar and related works, description of the method used in this paper, result of our method, discussions of issues

[1] https://entertainment.thaiza.com, accessed on 6 November 2019.
[2] https://people.mthai.com/starthai, accessed on 6 November 2019.
[3] http://www.siamdara.com/profile/thais, accessed on 6 November 2019.

encountered while we implement the method, and conclusion and future works.

II. Literature Review

Unstructured text is usually user-generated content that cannot be easily inserted into conventional databases.

Information Retrieval (IR) is query-based acquisition of a piece of information from a large collection of data. Using a search engine is basically an information retrieval. A main difference from IE is IR returns the result in natural language that humans can understand, while IE returns the result in a machine-understandable form for further processing.

Information Extraction (IE) serves as one of the core operations of text mining. The work of IE aims at recognizing specific types of information from objects of interests, events or relationships directly from text. Examples of the research attempts and techniques used are reviewed in this section. Reviewed works are grouped as lexicon-based works, free-text works, and personal information extraction works.

In this section we review lexicon-based works. Many information extraction systems and studies rely on a certain corpus and are in medical or biological fields. Liu et al. [1] developed AZDrugMiner to extract Adverse Drug-Event (ADE) from online patient forums as post-marketing drug surveillance is deemed a highly critical component of drug safety. AZDrugminer aimed to overcome inadequacies in existing system by extracting information from social media sites such as DailyStrength and PatientsLikeMe. Instead of lexicon-based extraction system like others, AZDrugMiner used machine learning to determine relationship, although many rule-based systems were also used. Likewise, Xu et al. [2] proposed a data-driven information extraction system for Chinese electronic medical records. This work used a Chinese medical term lexica and cross-domain dictionary to improve recall on named entities. The lexica were enriched further by pattern iteration. These systems had high precision but reliance on existing programs and infrastructure render this approach unsuitable for data that had little existing infrastructure.

Non-medical examples are work by Sasali et al. [3] who evaluated information extraction techniques using Malay documents such as magazines, novels, Quran, hadiths and other. The techniques used were lookup list, morphological rules (with noun affixes and verb/objective/noun affixes) and Rayner's Rule. It was found that morphological rules with verb/adjective/noun affixes was the most effective but further lexicon build-up was recommended.

In this section, Information extraction from open-text corpus include work by Sharma and De Choudhury [4] who developed a system that extracted nutritional information of food and ingestion content in Instagram. In this system, the researcher manually made a list of "canonical food names" or words that could query food-related Instagram posts. The USDA National Nutrient Database for Standard Reference database was used as a reference. The researcher processed the list of tags in each post and matched it with the list of canonical food. Also, there was a second method that compared the tag to the food descriptors in USDA. This method was reliable and accurate (89%) for food of moderate content. This work relied heavily on English language posts. On the other hand, Li et al. [5] created a web information retrieval system to extract news information from Baidu through analysis of URL of search results and the DOM structure of the web page. While it achieves high extraction and matching rate for open-text document, the search engine is limited to Baidu.

Works related with personal information extraction include that of Chen et al. [6] which proposed a hierarchical system that could extract personal information from resume PDF files. This work used conditional random fields and supporting vector machine to segment resumes into blocks and extract information. Li et al. [7] suggested a user profile extraction (Education, job and spouse) from Twitter. This approach requires only weak supervision but was limited to Twitter and job recall and precision were low as mentioning of job places in Twitter did not always mean that person was employed in such places.

Chen et al. [8] who introduced a robust web personal name information extraction system integrated with a heterogeneous attribute extraction and disambiguation system that extract features from multiple sources without supervision while achieving excellent precision. However, this approach has a relatively lower recall rate as each integrated AE approach covers only a percentage of the heterogeneous texts. Aboaoga and Ab Aziz [9] used rule-based approach for Arabic person name extraction. This approach used Introductory Words Person List and BAMA to recognize personal names. This study experimented on newspaper (sports, economy and politics) and found that named entity type, corpus size, number of keywords and rules, stop-words detection and morphological analyzer had impact on performance. It also found that this approached worked best on sports content.

There was one work that use statistical feature. Qu and Lu [10] investigated multi-translatable out-of-vocabulary terms which had received little attention and proposed a combined method that saw the use of statistical feature extraction, an artificial neural network combined with backward feature selection, and evolutionary parameter optimization.

Majority of the aforementioned works used corpus which was a finite, predictable set of data. Although there were some works on free-text acquired from a

search engine, those were primarily focused on other languages, which already had some infrastructure. The system proposed by this project mainly focuses on Thai-language free-text data extraction.

One study that described difficulties in Thai language named entity extraction is one by Imsombut and Sirikayon [11] on Thai tourism ontology, because agro attraction such as national parks had long word names, while shopping stores had larger variation of word names, in contrast with cultural attractions like temples and people's monuments.

The method proposed in this study uses a set of manually-made rules to automatically extract pieces of information related with date of birth, height and weight as "personal information candidates" from websites. Then, association rule measures and statistical rules which are co-occurrence, support, confidence, lift and conviction are used to select the most likely personal information candidate. Furthermore, weight of celebrities is estimated by using available height information and two sets of BMI values. Further details are specified in the following chapter.

III. Materials and Methods

This study aims to retrieve personal information birthday, height and weight from unstructured, Thai-language text such as website snippets. The experiment uses names of Thai celebrities as input. We also develop a weight estimation system for celebrities whose weight information is not obtained by the pattern-matching system, which estimate weight based on height and two sets of BMI values. The method consists of 4 parts: web crawling, rule-based pattern matching, selection by statistical method, and weight estimation using height and BMI index. The diagram is shown in Fig. 1 below:

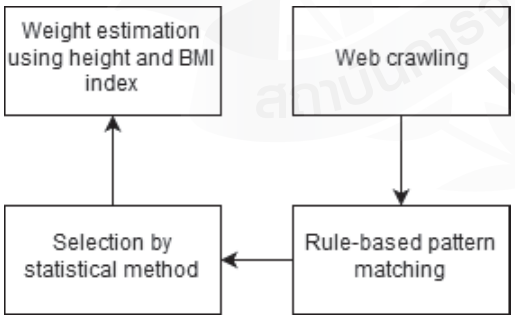


Fig. 1. Diagram of the information extraction process.

A. Web crawling

Web snippets related with Thai celebrities are the main source of information and acquired by using Google Custom Search API[4]. According to [10], the snippet limit should be 100 to ensure maximum coverage with acceptable noise. Example of the obtained snippets is shown in Fig. 2.

Fig. 2. Example of a web snippet from Google API.

A database is made to keep snippets for further use as shown in Fig. 3. In the table, title, URL, snippet, and the names of celebrities used to search for the snippet are stored. Each entry is given IDs.

SID	Title	Snippet	Web	Searcl
19...	บันเทิง - 'ลน' สวมวิญญาณ...	29 ก.ย. 2559 - เบรกความ...	www.naewna.com/enter...	รถบัส
19...	ข่าว'น้องรถบัส'ภูมิใจเป็นนัก...	14 ม.ค. 2560 - รถบัส : ส...	https://entertainment.ka...	รถบัส
19...	มาเรีย เข้าถึงอารมณ์เด็กเจ...	28 ก.ย. 2559 - กำลังเข้มข้...	https://www.siamzone.c...	รถบัส
19...	พูดคุยกับคู่ขี้ต่างวัย 'ฮัท จิร...	15 ม.ค. 2561 - 'ฮัท-จิรวิช...	www.trueplookpanya.co...	รถบัส
19...	Ladprao General Hospita...	... 2560" ณ ชั้นใต้ดิน ศูนย์...	www.ladpraohospital.co...	รถบัส
19...	Download รถบัส ด.ช.ภคพ...	Result for: รถบัส ด.ช.ภค...	www.aslady.de/video/รถ...	รถบัส
19...	chanyamcclory - Instagr...	เข้ากัน ที่สำคัญหน้าตาเหมื...	https://deskgram.org/ch...	รถบัส
19...	โป๊ป-มิว ขวนแฟนคลับผู้โช...	18 ม.ค. 2561 - ... โป๊ป-ธ...	www.becmultimedia.com...	รถบัส
19...	'ลน-แยม'นำทีมนักแสดง'ดว...	11 ก.ย. 2559 - ทิพย์รดา ไ...	m.innnews.co.th/mobile/...	รถบัส
19...	บ้านเมือง - คุจิ้นต่างวัย 'อั...	20 ธ.ค. 2560 - นานมัน 1 ...	www.banmuang.co.th/ne...	รถบัส
19...	บันเทิง - 'ลน' อึ้ง! 'มาเรีย' เ...	29 พ.ย. 2559 - น้ำตามาเร็...	www.naewna.com/enter...	รถบัส
19...	ดูดวงใจพิสุทธิ์ (จด) วันที่ 1...	14 ต.ค. 2559 - ... ลูกหมี ...	lakorn.guchill.com ? ละคร...	รถบัส
19...	หวนคืนบัลลังก์คู่จิ้นในตำน...	27 พ.ย. 2560 - ภคพล ต้...	https://www.tvpoolonline...	รถบัส

Fig. 3. Example of the snippet table.

B. Rule-based pattern matching

In this section, date of birth, height and weight of the celebrities are extracted using pattern matching approach, detail of each part shall be explained below. Regular expressions are used to extract "candidates" or possible words. Hand-crafted regular expression rules are used.

The pattern matching locates the input (celebrity names) in the snippet and then find a certain group of keywords that matches one of the rules. Then a personal information candidate is extracted based on such rules. If there is no matching, the next rule will be selected, and repeats until a match is found, or the rules are exhausted. Diagram of the pattern matching approach is shown in Fig. 4 below.

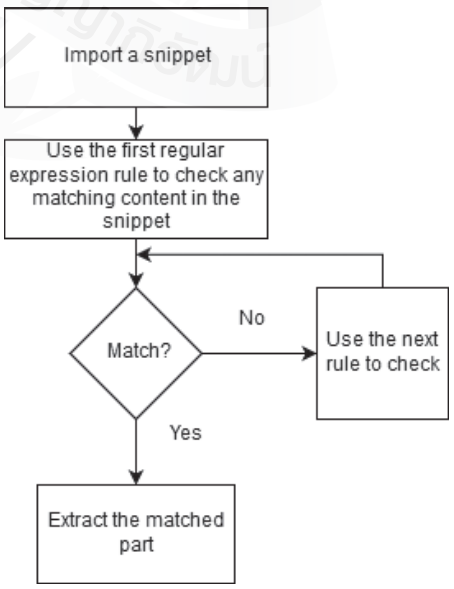


Fig. 4. Pattern matching diagram.

[4]https://developers.google.com/custom-search, accessed on 6 November 2019.

The first group of personal information candidates to be extracted is date of birth. Keywords such as "วันเกิด" or "เกิดวันที่" and "พ.ศ." (Buddhist Era) are used to locate and extract the candidate, as shown in Table I.

TABLE I
EXAMPLE OF EXTRACTED BIRTHDAY INFORMATION

Named entity	แอน อังคณา ทิมดี
Snippet	ประวัติ แอน อังคณา ทิมดี เกิดเมื่อวันที่ 9 กุมภาพันธ์ 2507 เป็นนักร้อง นักแสดง ผลงานละคร กุหลาบเหนือเมฆ และภาพยนต์ เรื่อง เขี้ยวอาฆาต.
Matching rules	%s.*?เกิด.*?([0-9]?[0-9]).?(\S+).*?([0-9][0-9][0-9][0-9])
Matched content	9 กุมภาพันธ์ 2507

As seen in Table I, birthday candidate of Angkana Timdee is 9 February 1964 as part of the snippet matches with one of the rules.

The second group of personal information candidates is height. In this part, the pattern matching approach would look for any 3-digit numbers near the keywords in the snippets, as it is improbable for an adult celebrity to have height less than 100cm unless they have medical issue (dwarfism etc.), before extracting it. Examples are shown in Table II.

TABLE II
EXAMPLE OF HEIGHT EXTRACTION

Named entity	นิศาชล ต้วมสูงเนิน เม
Snippet	3 ธ.ค. 2016 ... กลับมาพบกันอีกครั้ง กับ คอลัมน์ สวัสดีแคมปัส สัปดาห์พี่แคมปัส มีนางเอกช่อง 3 มาแนะนำ น่า รักสดใส สมวัย แถมดีกรีความสวยไม่ธรรมดา เราไปรู้จักเธอพร้อมๆ กัน. ชื่อ-นามสกุล : นิศาชล ต้วมสูงเนิน นิกเนม : เม วันเกิด : 24 เมษายน 2536 อายุ : 23 ปี พี่น้อง : มีน้องสาว 1 คน น้ำหนัก : 48 กิโลกรัม ส่วนสูง : 167 เซนติเมตร บ้านเกิด : อุดรธานี
Matching rules	ส่วนสูง.*:? (1\d{2})
Matched content	167

The third personal information candidate group is weight. The pattern matching approach looks for a keyword such as "น้ำหนัก" and then select any number behind it as shown in Table III.

TABLE III
EXAMPLE OF WEIGHT EXTRACTION

Title	รัศมีแข ฟ้าเกื้อล้น | GMM52 NEWFACE
Snippet	รัศมีแข ฟ้าเกื้อล้น. ชื่อเล่น. รัศมีแข. เพศ. ชาย. อายุ. 31 ปี 5 เดือน. น้ำหนัก. 89 กก. ส่วนสูง. 184 ซม. สัดส่วน. -. สีตา. ดำ. สีผม. ดำ. แบบผม. สั้น. ภาษา. ไทย/อังกฤษ/สวีเดน. รองเท้า. 9. ความสามารถพิเศษ.
URL	http://gmm.52com/newface/RDM8050
Matched content	89

In this section, 3 groups of personal information candidate groups are extracted. Due to the large number of candidates, statistical methods are used to select the correct personal information candidate which will be explained in the next section.

C. Selection by statistics

We use five statistical features for selecting the possibly correct personal information candidate, which are support, co-occurrence frequency, confidence, lift, and conviction. Each of the described statistical features is explained below.

1) Support

Support is the number of times a name of celebrity and a personal information candidate appear together divided by the total number of snippets. The formula is:

Where a is actor's name and b = information candidate $c_1 \ldots c_n$

$$Supp(a \to b) = \frac{f(a \cap b)}{N} \quad (1)$$

2) Co-occurrence frequency

Co-occurrence frequency is a number of times a celebrity name and a personal information candidate appears together.

Where a = actor's name and b = information candidate $c_1 \ldots c_n$

$$Co\text{-}occurrence\ (a \to b) = a \cap b \quad (2)$$

3) Confidence

Confidence is calculated by comparing the support of a celebrity name and a personal information candidate to support of the celebrity name.

Where a = actor's name and b = information candidate $c_1 \ldots c_n$

$$Conf(a \to b) = \frac{Support(a \cap b)}{Support(a)} \quad (3)$$

4) Lift

To get lift of a personal information candidate, divide the support of a celebrity name to the candidate by multiplication of individual support of celebrity and individual support of the candidate.

Where a = actor's name and b = information candidates $c_1 \ldots c_n$

$$lift(a \to b) = \frac{Support(a \cap b)}{Support(a)\,Support(b)} \quad (4)$$

5) Conviction

Conviction is division of support of the celebrity name by support of named entity that does not co-occur with candidate, and then support of candidate is compared against support of candidate without named entity.

Where a = actor's name and b = information candidates $c_1 \ldots c_n$

$$Conv(a \rightarrow b) = \frac{Support(a)}{Support(a \cap \neg b)} \quad (5)$$

D. Weight estimation using height and BMI index

Due to unavailability of weight information, an estimation method based on BMI index is proposed. BMI index is an indirect measurement of body fatness. The BMI index of less than 18.5 is considered underweight, and more than 24.9 is considered overweight. It is calculated by dividing the weight (in kilograms) with a squared height (in meters)

As celebrities tend to be fitter (male), or skinnier (female) than the average population, it is assumed that female BMI index should be slightly under average to maintain the slim body shape. On the other hand, muscular male celebrities should have BMI slightly over average due to muscle weight. According to Kate Stinchfield of Health.com[5], Angelina Jolie has an estimated BMI of only 17.9, while Arnold Schwarzenegger is estimated to have BMI of 30.8.

From our previous experiment, we establish 2 sets of BMI values for weight estimation. The first set of BMI values used numbers 18 and 22 (slightly underweight for female and near the upper healthy limit for male) as the BMI values for female and male celebrities. Then, the second BMI value for male and female celebrity weight estimation is calculated from average BMI of celebrities with known weight information, which are 16.7 and 20.2 respectively. Details of the BMI formula are as follows:

$$BMI_{(f,m)} = \frac{body\ weight(kg)}{body\ height(m)^2} \quad (6)$$

Where f = (18,16.7), m = (22,20.2)

In addition to our modified BMI formula, we also use a formula based on Martin Berkhan's ripped BMI (fat level at 5-6%) calculation per the following formula.

$$BMI_{ripped} = Height(cm) - 100 \quad (7)$$

Berkan's formula, however, is developed for athletes.

IV. Experiment and Result

A. Web crawling

We used 317 celebrity names presented on MThai.com/starthai, as input. Snippets were obtained by using Google Custom Search API, using the celebrity name as keywords for searching. In total we obtained 22,484 snippets, an average of 70.92 snippets per one celebrity. Some celebrities were relatively obscure and resulted in less than 100 snippets being attributed to them. Additionally, a master student was hired to manually collect personal information to make a baseline for comparison.

B. Pattern matching

The recall number of possible candidates for birthday extracted is 458, height is 98 and weight is 119.

C. Selection by statistical methods

This section includes 3 accuracy graphs for the 3 celebrity candidate groups, and one graph for BMI estimation. Fig. 5 Shows a graph for date of birth.

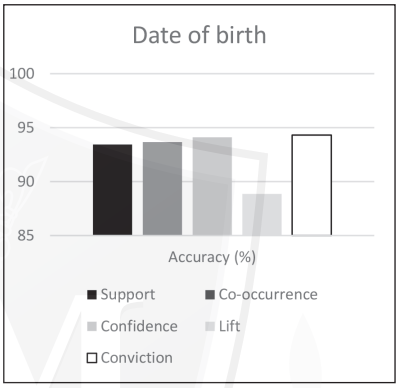


Fig. 5. Accuracy of date of birth features.

As seen in Fig. 5, conviction has the highest accuracy at 94.32%, followed by confidence, co-occurrence, support and lift respectively.

Most of the incorrect result were from misattribution as some celebrities did not have birthday information. Another problem was some snippets included a list of celebrities and their birthdays. As a result, the system happened to pick the birthday of someone other than the intended entity. Misspelling in words such as "พ.ศ." or "กรกฎาคม" also affected accuracy. Also from observation, candidates tend to be misattributed if multiple birthdays and actors are included in the same snippet.

If a snippet includes a list of celebrities and their birthdays, which still counts as co-occurrence despite misattribution. For this reason, some candidates have identical confidence. One Thai celebrity, Alice Tsoi, has relatively obscure birthday information and result in multiple contesting candidates (1 and 4 June 1991).

The next part is result from height candidate group. Fig. 6 below shows accuracy of height candidate features.

[5]https://www.health.com/health/gallery/0,,20460621,00.html, accessed on 6 November 2019.

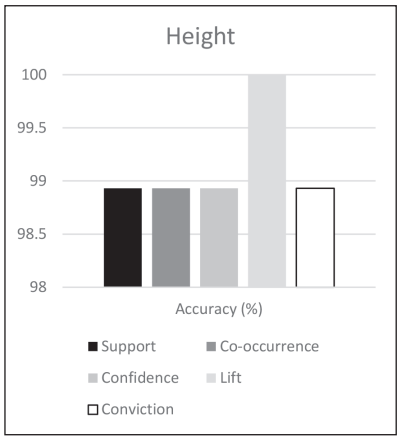


Fig. 6. Accuracy of height candidates.

According to Fig. 6, accuracy for most features are equal at 98.93% except lift, which each correct candidate has higher than all other candidates.

On height, assessment of candidates showed that out of 98 height candidates, 4 candidates do not have co-occurrence with the named entities. Gee-Thanyachon does not co-occur with her correct 170cm candidate because the candidate instead appears with her old name: Rattakorn. Similarly, Gale-Waethaka has height information under her old name Wipawee. Another is due to misspelling of the nickname in the snippet. The highest number of co-occurrence is 9, with the celebrity name Mak Parin.

Out of 94 celebrities with collected height candidates, only 4 celebrities have two height candidates as shown in Fig. 7 below.

Heightcan	ActorID	SupportCooc	Confidence	Lift	Conviction
160	137	0.0000444741	0.0238096	35.6905	1.02439
162	137	0.0000444741	0.0238096	89.2262	1.02439
165	189	0.0000444741	0.0140845	17.5939	1.01428
167	189	0	0	0	1
165	311	0.0000889482	0.0136054	16.9955	1.01379
167	311	0.0000444741	0.00680272	12.7466	1.00685
168	313	0.000266845	0.1	149.9	1.11111
174	313	0.0000889482	0.0333333	53.5357	1.03448

Fig. 7. Features of all ambiguous height candidates.

As seen on Fig. 7, the candidates for the celebrity no. 137 have similar values in all but Lift. For the celebrity no.189 (Gale-Waethaka), the 167 cm candidate does not have co-occurrence with the named entity due to name changes as mentioned. Named entities number 311 and 313 each has one candidate with more confidence, lift and conviction then the other, and is the correct candidate.

The next personal information candidate group to be evaluated is weight. Result is shown in the following Fig. 8.

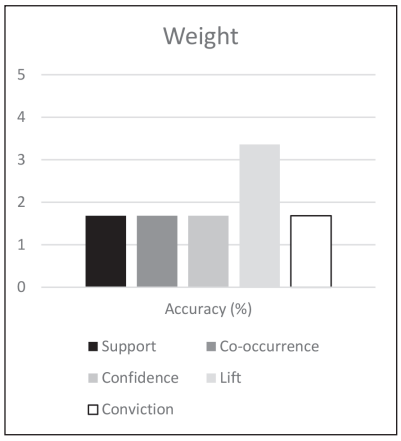


Fig. 8. Accuracy of weight candidates' features.

According to Fig. 8, accuracy is very low compared to date of birth and height, with maximum accuracy being only 3.36% for lift.

While 100 celebrities have weight attributed to them, only 44 attributed weight values are deemed plausible by manual inspection. The remaining 66 celebrities have attributed weight of less than 40 kilograms which is impossible for a well-looking actor/actress, except one who is a child actor).

The next part is weight estimation using height, 2 sets of BMIs (18 and 16.7 for female and 22 and 23.2 for male celebrities) and Berkan's fomula. As only 94 celebrities have height attributed, weight estimation is calculated only for them. Result of weight estimation is shown in Fig. 9.

FName	LName	NickNa...	Gender	Weight	BMIEst
รัศมีแข	ฟ้าเกื้อลัน	รัศมีแข	M	89	(NULL)
คณัสนันท์	นักตะเข่	มาร์ค	M	69	71.28
นราวิชญ์	จิตรบรรจง	เขม	M	68	68.9238
วัทธิกร	เพิ่มทรัพย์หิรัญ	เกี๊ยก	M	68	71.28
ธนวรรธน์	วรรธนะภูติ	โป่ป	M	65	67.375
อรรคพันธ์	นะมาตร์	อ่อม	M	73	72.0742
พิชญะ	นิธิไพศาลกุล	กอล์ฟ	M	53	63.58
กิจเกษม	แมคแฟดเดน	เจมส์	M	58	71.28
ศตวรรษ	เศรษฐกร	เต๊ะ	M	60	66.6072
ปกรณ์	ฉัตรบริรักษ์	บอย	M	63	71.28

Fig. 9. Complete weight information of male celebrities at BMI = 22 (gender, obtained weight, BMI estimation).

As visible in Fig. 9, the estimated weight is higher for most male celebrities except 3 male celebrities who weigh more than the BMI estimation, out of 20 male celebrities with attributed weight. One male celebrity does not have BMI calculation due to absence of height information.

An example of obtained and estimated weight of female celebrities is shown below:

FName	LName	NickName		Weight	BMIEst
ณปภา	ตันตระกูล	แพท	F	43	52.9891
ตฤณญา	มอร์สัน	แคท	F	51	56.2096
พิชชาภา	พันธุมจินดา	แพร์	F	49	54.91
ณัฐวรา	วงศ์วาสนา	มิ้นท์	F	45	51.7275
พรรัมภา	สุขใต้พึ่ง	แพร	F	49	56.8651
กัญจน์ณิช	กิตติพรภาณุวงศ์	แก้วใส	F	45	51.7275
คามิลล่า	กิตติวัฒน์	มิลลี่	F	48	51.1024
พีชญา	วัฒนามนตรี	มีน	F	47	51.7275
เดบาร่า	ซี	เด็บบี้	F	45	49.8636
แอน	ทองประสม	แอน	F	49	50.4811
เวธกา	ศิริมณีวัฒนา	เกล	F	47	51.7275
อัมราภัสร์	จุลกะเศียน	มิ้ม	F	51	56.8651

Fig. 10. Example of weight information of female celebrities at BMI = 18 (Obtained weight and BMI estimation).

On the other hand as seen in Fig. 10, female celebrities' obtained weight is generally lighter than the BMI estimation. It is notable that no female celebrity is heavier than the BMI estimation.

Then, the second set of BMI values (16.7 for female and 20.2 for male celebrities) is used. Fig. 11 below is a complete result of modified estimation for male celebrities.

FName	LName	NickName	...	Weight	BMIEstMod
คณัสนันท์	นักตะเข่	มาร์ค	M	69	65.448
นราวิชญ์	จิตรบรรจง	เขม	M	68	63.2846
วัทธิกร	เพิ่มทรัพย์หิรัญ	เกี๊ยก	M	68	65.448
ธนวรรธน์	วรรธนะภูติ	โป๊ป	M	65	61.8625
อรรคพันธ์	นะมาตร์	อ๋อม	M	73	66.1772
พิชญะ	นิธิไพศาลกุล	กอล์ฟ	M	53	58.378
กิจเกษม	แมคแฟดเดน	เจมส์	M	58	65.448
ศตวรรษ	เศรษฐกร	เต๊ะ	M	60	61.1575
ปกรณ์	ฉัตรบริรักษ์	บอย	M	63	65.448
เศรษฐพงศ์	เพียงพอ	เต่า	M	65	64.0017

Fig. 11. Example of modified (BMI = 20.2) weight estimation (the rightmost column) compared with the obtained weight for male celebrities.

As seen in Fig. 11, 2 male celebrities have BMI estimation being within 1 kilogram of the obtained weight.

The following part is the modified BMI estimation result for female celebrities.

FName	LName	NickN...	...	...	BMIEstMod
ณปภา	ตันตระกูล	แพท	F	43	46.5746
ตฤณญา	มอร์สัน	แคท	F	51	49.4053
พิชชาภา	พันธุมจินดา	แพร์	F	49	48.263
ณัฐวรา	วงศ์วาสนา	มิ้นท์	F	45	45.4658
พรรัมภา	สุขใต้พึ่ง	แพร	F	49	49.9814
กัญจน์ณิช	กิตติพรภาณุวงศ์	แก้วใส	F	45	45.4658
คามิลล่า	กิตติวัฒน์	มิลลี่	F	48	44.9163
พีชญา	วัฒนามนตรี	มีน	F	47	45.4658
เดบาร่า	ซี	เด็บบี้	F	45	43.8275
แอน	ทองประสม	แอน	F	49	44.3702
เวธกา	ศิริมณีวัฒนา	เกล	F	47	45.4658

Fig. 12. Example of modified BMI estimation for female celebrities.

Out of 33 female celebrities with attributed height and weight, 10 female celebrities have BMI estimation being within 1 kilogram of the obtained weight.

Then, weight estimation based on Berkhan's formula is used. It was found that no celebrity with attributed weight meets the "ripped" weight or higher, suggesting lower muscle mass and likely poor suitability of the formula outside of athletic circles. Example of ripped weight estimation is shown in Fig. 13 below:

FName	LName	NickName	G...	Weight	BMIEstlean
ณปภา	ตันตระกูล	แพท	F	43	67
ตฤณญา	มอร์สัน	แคท	F	51	72
คณัสนันท์	นักตะเข่	มาร์ค	M	69	80
พิชชาภา	พันธุมจินดา	แพร์	F	49	70
นราวิชญ์	จิตรบรรจง	เขม	M	68	77
วัทธิกร	เพิ่มทรัพย์หิรัญ	เกี๊ยก	M	68	80
ณัฐวรา	วงศ์วาสนา	มิ้นท์	F	45	65
ธนวรรธน์	วรรธนะภูติ	โป๊ป	M	65	75
พรรัมภา	สุขใต้พึ่ง	แพร	F	49	73
กัญจน์ณิช	กิตติพรภาณุวงศ์	แก้วใส	F	45	65

Fig. 13. Example of ripped weight estimation (the rightmost column) compared with the obtained weight.

V. Discussion

Although conviction is the most accurate feature, at least one record (Ann Thitima's birthday), has 8 February 1988 and 4 July 1979 as possible candidates. Although correct, the 4 July candidate has conviction of only 1.09302 compared with the erroneous 8 February 1988 (actually another celebrity's birthday on the same page) that has conviction of 1.175. In another case, the 8 February 1988 is also a candidate for Bow Sunita (who was actually born in 1975), and has higher lift and conviction than the correct candidate. Despite this, high confidence (0.0434782 compared to only 0.101449) ensure that the 1975 candidate is correct.

In addition, misspelling or use of old names can lead to incorrect candidate extraction. In our experiment, 2 celebrities had their correct weight information under their old names and as a result were not selected.

VI. Conclusion

In conclusion, we have developed a system to automatically extract personal information candidates using novel rules. Moreover, association rule measures and statistical methods were used to select the possibly correct information. Furthermore, a modified BMI-based weight estimation method is proposed to overcome absence of data weight problem. We tested the methods with Thai celebrity from MThai.com and our method gained more information than MThai.

In future works, we plan to employ machine learning to find other information such as work/romantic relationship between actors, films, or production crewmembers.

Authors' Contribution

The first author conducted the experiment and drafted the manuscript. The last author guided the experiment and co-drafted the manuscript. The first author and the last author contributed 85% for this research article. The second, third, and fourth authors reviewed the manuscript and provided feedback.

References

[1] X. Liu and H. Chen, "AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums," International conference on smart health, pp. 134-150, 2013.

[2] D. Xu, Meizhuo Z, and et al., "Data-driven information extraction from Chinese electronic medical records," PloS one, vol. 10, no. 8, pp. 1-18, 2015.

[3] S. S. Sazali, N. A. Rahman, and Z. A. Bakar, "Information extraction: Evaluating named entity recognition from classical Malay documents," in 2016 Third International Conference on Information Retrieval and Knowledge Management (CAMP), 2016, pp. 48-53.

[4] M. De Choudhury, S. Sharma, and E. Kiciman, "Characterizing dietary choices, nutrition, and language in food deserts via social media," in Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing, 2016, pp. 1157-1170.

[5] J. Li, G. Jiang, A. Xu, and Y. Wang, "The Automatic Extraction of Web Information Based on Regular Expression.," JSW, vol. 12, pp. 180-188, 2017.

[6] Y. Chen, J. Zhou, and M. Guo, "A context-aware search system for internet of things based on hierarchical context model," Telecommunication Systems, vol. 62, no. 1, pp. 77-91, Mar. 2016.

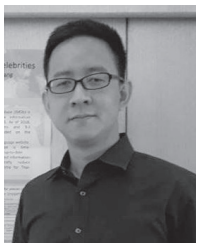
[7] J. Li, A. Ritter, and E. Hovy, "Weakly supervised user profile extraction from twitter," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 165-174, 2014.

[8] Y. Chen, S. Y. M. Lee, and C.-R. Huang, "A robust web personal name information extraction system," Expert Systems with Applications, vol. 39, no. 3, pp. 2690-2699, 2012.

[9] M. Aboaoga and M. J. Ab Aziz, "Arabic person names recognition by using a rule based approach," Journal of Computer Science, vol. 9, no. 7, pp. 922, 2013.

[10] J. Qu and Y. Lu, "Automatic identification and multi-translatable translation of vocabulary terms with a combined approach," in 2016 Eighth International Conference on Advanced Computational Intelligence (ICACI), 2016, pp. 342-348.

[11] A. Imsombut and C. Sirikayon, "An alternative technique for populating thai tourism ontology from texts based on machine learning," in 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS), 2016, pp. 1-4.



Chinorot Wangtragulsang was born in 1988 in Bangkok, has a Bachelor's degree in Computer Science from Sirindhorn International Institute of Technology (SIIT), Thammasat University (graduated in 2012). After graduation, he was drafted into the Royal Thai Army in 2014 and discharged in the following year with outstanding contribution to the army. After that, he works as a teaching assistant at SIIT and also takes up other part-time professions such as a freelance translator (since 2016), computer instructor at SIAM Computer (since 2019) and a military interpreter working for various Army and Police units (since 2017).

Nattakarn Phaphoom is currently a lecturer at Panyapiwat Institute of Management, Thailand. She obtained a Ph.D. from Free University of Bozen-Bolzano, and double master's degree in software engineering from Blekinge Institute of Technology, Sweden, and Free University of Bolzano-Bozen, Italy. Prior to studying the master's degree, she worked for 3 years for IBM Solutions Delivery Co., Ltd., Thailand. Her research interests include emerging technologies, innovation adoption, cloud computing, data analytics, and agile methods.

Phannachet Na Lamphun received the Master of Science (Computer Engineering), Polytechnic institute of New York University, New York, USA and Doctor of Engineering (Information and Communications Technologies), Asian Institute of Technology.

Past work experience is at IndexInternational Group Co.,Ltd. as a system engineer for Project Management Consultation of MRTA Blue Line, Dustfree Road Projects, and Water Forecast and Flood Alarm Project.

Pisit Charnkietkong has the M.Eng. and D.Eng. in Electrical & Computer Engineering from Yokohama National University, Japan.

Currently, he is a dean of Panyapiwat Institute of Management. In the past, he was a dean of Faculty of Information Technology, Rangsit University, a head of Electrical Engineering department, Faculty of Engineering, Sripatum University.

Jian Qu is a full-time lecturer at the Faculty of Engineering and Technology, Panyapiwat Institute of Management. He received Ph.D. in information science with Outstanding Performance award from Japan Advanced Institute of Science and Technology, Japan, in 2013. He received B.B.A with Summa Cum Laude honors from Institute of International Studies of Ramkhamhaeng University, Thailand, in 2006, and M.S.I.T from Sirindhorn International Institute of Technology, Thammasat University, Thailand, in 2010. He was a program director and lecturer for School of information technology, SIU, Thailand from 2013 to 2017, and secretary to CEO and international marketing manager at Eason Paint PCL from 2007 to 2010. His research interests are natural language processing, intelligent algorithms, machine learning, machine translation, information retrieval and image processing.