Novel Phonetic Name Matching Algorithm with a Statistical Ontology for Analysing Names Given in Accordance with Thai Astrology

Chakkrit Snae and Michael Brückner
Faculty of Science, Naresuan University, Phitsanulok, Thailand

chakkrits@nu.ac.th; michaelb@nu.ac.th

Abstract

Since antiquity names have been very important to people. Naming from the past to the present has been continuously developed and has evolved into a variety of patterns. Each pattern has its own rules depending on local belief and language that has been developed until the present. In many cultures naming is not only important because every individual needs to have a name but have helpful names or names with a good sound. The basic goal of naming in Thai society is to provide a good fortune and progress of living. Most Thai parents try to choose names they feel will bring good luck to their offspring and to the family. The choice of appropriate names is based on old rules of Thai astrology according to weekday of birth, and the rules of available letters can influence the destiny of the individuals as is described in Thai astrology, since it uses the day of birth as an input. Thais can change their own given names as often as they want in order to achieve a good fortune. The current web based systems for Thai names are static web pages and cannot deal with the problem of helping change a name to a good name with similar sound.

In this research, a web-based system with a novel name matching algorithm for analysing Thai names is proposed, which takes into account the Thai astrology and uses a statistical ontology to check and evaluate how suitable names in the cultural environment with respect to sound and the persons' fortune are. The system and the algorithm have been implemented to assess Thai naming habits and the development in naming conventions over the past 20 years. The analysis concentrates on how which names have been adopted as “good names”, how much they follow the rules according to Thai astrology and whether they contain letters out of the so-called misfortune attribute set or not. After a name has been found to be of low value to the individual or to contain letters from the misfortune attribute, the system of Thai astrology naming will help to change names. A new composite name matching technique called Metasound (a combination of the Soundex and Metaphone algorithms) has been implemented and is used for finding name variants (spelling and phonetic variations). Therefore, Metasound has been developed based on commonplace rules of Thai pronunciation for matching words that sound and are spelled alike. The algorithm reduces the Thai alphabet to eight consonant sounds:/ k/ or /kh//K,/b/ or /p/ B, /d/, /t/ or /th/ D, //ŋ/ NG, /n/ N, /m/ M, /j/ Y and /w/ W. With the help of this new algorithm it is possible to offer a web based service for changing a given name leading to a good fortune according to Thai belief and with a melodious sound.

Keywords: Thai naming system, statistical ontology, rule based system, phonetic algorithm

Introduction

Names are used for referring to people, places, things, and even ideas or concepts, and in many cases for identifying them. Names serve as labels of categories or classes as well as individual items or instances. They are properties of individuals, which are of major importance in most communities and this also the case for Thailand. The way by which Thai parents get their children names can vary, e.g. naming by monks or grandparents. The traditional naming of Thai children from the past to the present has been continuously developed into a variety of patterns. Each pattern has its own rules with regional variations and depending on the belief developed during the centuries. The basic goal of naming in Thai society is to provide a good fortune and progress during life, and this is done by choosing given names (or first names) carefully. Most first names have a meaning. The naming methodology used in this research is the most widely used naming system, which uses Thai astrology according to the weekday of birth, unlike the Western astrology, which is based on the zodiac and the date of birth. Most Thai believe that the individual has a set of 8 attributes called “name of the angles” referred to in Thai astrology (Snae & Brueckner, 2006a), which influence each person’s livelihood, fortune, and so on. The attributes are called Servant, Age, Power, Honour, Property, Diligence, Patron, and Misfortune. Each attribute has it own letters that can be used for constructing good names.

Nowadays, Thai naming systems can be seen in the Internet but most of them are on static web pages keeping indexes of names (according to birthdates) and their meanings in a database. There are some disadvantages in the current systems: (1) only a small number of names are there in the databases, around 3,000 to 4,000 names, (2) no opportunity is provided to change old names to good ones similar to the old ones with a better meaning.

To tackle names and their variations phonetic algorithms have been used for a long time. A phonetic algorithm is an algorithm for indexing of words by their pronunciation. Most phonetic algorithms were developed for use with the English language; consequently, applying the rules to words in other languages might not give a meaningful result. They are necessarily complex algorithms with many rules and exceptions, because English spelling and pronunciation is complicated by historical changes in pronunciation and words borrowed from many languages. Among the algorithms in use are the Soundex, the Metaphone and the NYSIIS algorithms together with their numerous variations.

Soundex is a phonetic algorithm to enable retrieve of information from data processing systems. R. C. Russell developed the Soundex algorithm to processes data collected from the 1890 census. Known as the Russell Soundex algorithm numerous variants have been employed for genealogy studies and retrieval systems. The Soundex algorithm has been adapted to the Thai language by Lorchirachoonkul (1982) by taking into account the specific characteristics of the Thai language.

Metaphone is a phonetic algorithm for indexing words by their sound, when pronounced in English. The algorithm produces variable length keys as its output, as opposed to Soundex's fixed-length keys. Similar sounding words share the same keys.Metaphone was developed by Lawrence Philips as a response to deficiencies in the Soundex algorithm. It is more accurate than Soundex because it uses a larger set of rules for English pronunciation. Metaphone is available as a built-in operator in a number of systems, including later versions of PHP.

In 1970 the New York State Identification and Intelligence project headed by Robert L. Taft published the paper "Name Search Techniques". In this paper he compared Soundex with a new phonetic routine (NYSIIS) that was designed through rigorous empirical analysis.

The term ontology has been widely used in recent years in the field of Artificial Intelligence, computer and information science especially in domains such as, cooperative information systems, intelligent information integration, information retrieval and extraction, knowledge representation, and database management systems. Many different definitions of the term are proposed. One of the most widely quoted and well-known definition of ontology is Gruber's (1993): An ontology is an explicit specification of a conceptualization.

The concept of ontologies can be combined with statistics, in which case it is called statistical ontology and mostly used in data analysis, data mining or clustering of relational data and their display as statistical values (Denk, Froeschl, & Grossmann, 2002; Hert & Haas, 2003). Marchinonini, Haas, Plaisant, Shneiderman, and Hert (2003) developed a statistical ontology for finding relation of statistic and link concept of terms together by constructing definition, graph illustrating development. The ontology was used for constructing and illustrating support. Pasquier, Girardot, Jevardat de Fombelle, and Christen (2004) developed a tool called THEA (Tool for High-throughput Experiments Analysis) for analyzing huge experiment work. This tool used statistical ontology concept for constructing general meaning from basic knowledge clustering and retrieval. The data retrieval used word explanation to search knowledge in biology and used data mining for data and knowledge clustering which automatically outputted into tree pattern. In this research we use a statistical ontology for analyzing and checking good names.

Recently, Snae and Brueckner (2006a, 2006b) have pioneered to develop a dynamic online Thai naming system based on Thai astrology, which uses letters according to the day of birth to construct a melodious sounding name with a meaning and adopts the name matching algorithm LIG3 (Levenshtein, Index of Similarity Group (called ISG), and Guth) for finding similar names and variants (Snae, 2007). However, LIG3 seems to be more complex and time consuming because it contains many functions, e.g. a function to calculate the distance between two names using Levenshtein method (Levenshtein, 1965) and a function similarity measurement. A recently proposed web based Thai naming system uses various techniques, such as a rule base (Snae & Brueckner, 2006a), a hybrid name matching method (Snae & Brueckner, 2006b), an ontology of names (Snae, 2006), and a Thai name checking system using Thai astrology (Snae & Namahoot, 2007).

In this paper we propose an improved web-based Thai naming and name checking system that uses a novel phonetic algorithm MetaSound, the Thai astrology naming concept, and an appropriate ontology supporting the analysis and evaluation of good names in the cultural environment with respect to fortune and sound. The analysis is focussed on how which names have been adopted as “good names”, how much they did follow the concept of Thai astrology and whether they contain letters out of the misfortune attribute set or not. The proposed system can alternatively generate good names that are similar to old ones and is designed to find name variants (taking into account spelling as well as phonetic variations).

This paper is organized as follows: first we give a description of principle naming using Thai astrology for given names. Secondly, we explain how the ontology and the statistical ontology will be applied for the Thai name checking system. After that we present some basic characteristics of the name matching algorithms of our choice, and a new phonetic based algorithm, as well as the system for naming applied to Thai names. Then the results of Thai name checking and Thai astrology naming systems can be shown. The last section shows the conclusions of our study and further work which has to be performed.

Principle Naming using Thai Astrology for Given Names

The way of naming a Thai child can vary, for instance naming by monks or by the grandparents. Naming from the past to the present has been continuously developed and has evolved into a variety of patterns. Each pattern has its own rules depending on local belief and language that has been developed until the present. The basic goal of naming is to provide a good fortune and progress during life. Most first names have a meaning. Principal naming using Thai astrology is widely used since the time of the old kingdoms of Siam, and it involves the weekday of birth in order to construct the name. This is based on the belief that the individual has a set of 8 attributes called name of the angles referred to in Thai astrology. These attributes influence each person's livelihood, fortune, and so on. The attributes are called Servant, Age, Power, Honour, Property, Diligence, Patron, and Misfortune. Each attribute has a distinct set of letters that can be used for constructing names of melodious sound and good fortune.

In many cultures naming is not only important because every individual needs to have a name but to have helpful names or names with a good sound. Thai parents always try to choose names which they feel will bring good luck to their offsprings and to the family. The choice of appropriate names bases on the rules of available letters that can influence the destiny of the individuals as described in the following. Letters and days refer to Thai astrology. In that process the fortune depends on the day of birth and the related letters shown in Table 1. Note that Wednesday occurs twice. This is because counting of the Thai dates is different from the Western style. Thai people start a week on Sundays and a new day at 6 a.m. Thus naming using dates has to consider timing as well, especially on Wednesdays, which is the middle of the week and can be divided into daytime (from 06:00 to 17:59) and nighttime (18:00 to 05:59 the next day).

Starting with the weekday of birth here are a set of letters available, which refer to eight basic properties as follows:

- Servant: children, Husbands, wives, including people who we support within family
- Age: life, livelihood, including the way of living
- Power: destiny, honor, fame, position, including education and love
- Honor: asset, money, appliance fortune which can be gained in the future
- Property: assets that are inherited and still exist in the present including status of relatives
- Diligence: success from working, including creative and hard working
- Patron: Supporters, such as parents , teachers, bosses and helpers
- Misfortune: Evil, enemies, sins, including any obstacles

Table 1: Letters of the angels referred to in Thai astrology (Snae and Namahot, 2006)

Day	Servant	Age	Power	Honor	Property	Diligence	Patron	Misfortune
Sunday	All vowels	กขคฆง	จฉชฌญ	ฎฏฐฑฒ ณ	ดตถ ทธน	บปผฝพ ฟภม	ยรลว	ศษสหฬอ
Monday	กขคฆง	จฉชฌญ	ฎฏฐฑฒ ณ	ดตถ ทธน	บปผฝพ ฟภม	ยรลว	ศษสหฬอ	All vowels
Tuesday	จฉชฌญ	ฎฏฐฑฒ ณ	ดตถ ทธน	บปผฝพ ฟภม	ยรลว	ศษสหฬอ	All vowels	กขคฆง
Wednesday daytime	ฎฏฐฑฒ ณ	ดตถ ทธน	บปผฝพ ฟภม	ยรลว	ศษสหฬอ	All vowels	กขคฆง	จฉชฌญ
Wednesday Night	ยรลว	ศษสหฬอ	All vowels	กขคฆง	จฉชฌญ	ฎฏฐฑฒ ณ	ดตถ ทธน	บปผฝพ ฟภม
Thursday	บปผฝพ ฟภม	ยรลว	ศษสหฬอ	All vowels	กขคฆง	จฉชฌญ	ฎฏฐฑฒ ณ	ดตถ ทธน
Friday	ศษสหฬอ	All vowels	กขคฆง	จฉชฌญ	ฎฏฐฑฒ ณ	ดตถ ทธน	บปผฝพ ฟภม	ยรลว
Saturday	ดตถ ทธน	บปผฝพ ฟภม	ยรลว	ศษสหฬอ	All vowels	กขคฆง	จฉชฌญ	ฎฏฐฑฒ ณ

Consider the following example referring to Table 1: people who are born on a Sunday and need to add fortune in the attribute Servant need to use a set of vowels in naming process. For a boy the first letter must only be one of the letters in the attribute "Power" e.g. จ ฉ ช ฌ ญ. If they want the child to live a long life parents have to choose one of the letters in the attribute "Age". The same applies for the attribute "Property" if the offsprings should have a high status in their lives. However, the parents should not use letters out of the set of letters in "Misfortune", in the example ศ ษ ส ห ฬ อ, in order to avoid evil, enemies, sins or any other obstacles.

The past generations of Thais widely used letters in the "Power" set (attribute) for boys' names and letters in the "Honor" set for girls' names. Nowadays there is no limit in choosing letters for genders but letters from "Misfortune" attribute are generally avoided in the naming process.

Ontology and Statistical Ontology for Thai Name Checking System

An ontology is a branch of philosophy that deals with the science of what is, of the kinds and structures of objects, properties, events, processes and relations in every area of reality. The term ontology has been traditionally used in philosophy as a synonym for metaphysics, providing a systematic account of existence, and hence identifying, the subject of existence. However, ontology and metaphysics are far from being univocal and determinate in philosophical jargon, and important distinctions seem often enough to be marked by them. What one may call ontology is the attempt to say what entities exist (e.g. one's ontology is one's list of entities). Metaphysics, by contrast, is the attempt to say, of those entities, what they are (e.g. one's metaphysics is an explanatory theory about the nature of those entities).

In order to compute the goodness of names, the Thai name checking system will check each letter of a name matched in each attributes of Thai astrology. Users will enter a name and the day of birth, then the system will segment names into letters and count the number of letters (N), the day of birth will compare with each attribute of the Thai astrology table (Table 1) and arrange letters into attributes according to the day of birth and count the number of letters that match in each attribute (M(i)). The goodness of each attribute will be calculated and summed into the percentage of statistical goodness (G) of names, cf. (1). This G will exclude the calculation of misfortune attribute.

$$G = \sum_{i=1}^{n} \frac{M(i)*100}{N} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (1)$$

An example of name calculation using (1) can be illustrated as follows:

Users enter name “จักรกฤษณ์” (pronounced as chakkrit) and “Monday” as day of birth. Then letters in names จักรกฤษณ์ will be segmented and arranged into attributes of Thai astrology:

จ (ch) is in attribute Age

ก (k) is in attribute Servant

ก (k) is in attribute Servant

ร (r) is in attribute Diligence

ษ (s) is in attribute Patron

ณ (ch) is in attribute Power

As we can see, two letters in attribute Servant then the goodness of this attribute is (2/9)*100 = 22.22%

Age, Diligence, Patron, and Power attributes are represented by one letter each so that the goodness value of the four attributes is (1/9)*100 = 11.11%. Therefore the statistical goodness (G) of this name (chakkrit) is 22.22+11.11(* 4) = 66.67%.

The output of statistical goodness (G) of names using formula 1 will be used to divide ontological goodness values (Og) interval as follows:

If G> 79% then Og = 2

If 49%< G <= 79% then Og =1.5

If 20% < G <= 49% then Og = 1

If 0 % < G <= 19 % then Og = 0.5

If names do not contain a misfortune attribute then the ontological goodness values (O_m) is 1 otherwise O_m is -1.

After that, the system will calculate the reliability (R) of good names using a formula as follows:

$$R = \frac{O_g + O_m}{Max(O_g + O_m)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.... (2)$$

In our example above, Chakkrit, who was born on a Monday, has the statistical goodness (G) 66.67 % which leads to the ontological goodness value (Og) of 1.5. Also this name contains no misfortune attribute (O_m=1) therefore, the reliability (R) of this name is (1.5 + 1)/3 = 83.33%

Matching Algorithms for Name Similarity

When it is simply said that "two names are similar," the specific qualitative dimensions may be identified according to which the names are "close." Qualitatively, names may have similar structural tones, similar sounds, similar harmony; they may both describe similar meaning. Thus similarity depends upon our qualitative frame of references, for example, the names may be dissimilar in sounds, but similar in spelling or meaning.

Other qualitative dimensions are less immediately characterised quantitatively. For instance, having identified that "butter" and "batter" are similar in sound, we can note that the word "putter" is more similar to "butter" and the word "matter" is less similar. The absence of a standard measurement is inconvenient, but it is not necessarily problematic. In real applications an appropriate measurement can often be invented, although the resolving power of such a yardstick may be quite crude. For instance, in the case of similar-sounding words, perceptual experiments that provide some quantitative data indicating the relative frequency of confusion between various phonemes might be carried out.

More recently published name matching techniques are either of the composite or hybrid form (Snae & Diaz, 2002) and several novel hybrid algorithms have been developed for specific purposes. Snae (2007) distinguished and implemented four types of name matching algorithms as follows:

1) Spelling based algorithms rely on the assumption that the source and target names are strings, which differ because of one or more errors (insertion, deletion, substitution and transposition). The spelling of English names is usually arbitrary with no absolute rules that may be checked against a dictionary. Example algorithms of these types are “Guth” and “Levenshtein”.

- Guth algorithm. This type name is based on the approach due to Guth (Guth, 1976). The method is left to right sequence driven checking the three next letters for agreement, and it is essentially alphabetic but independent from language and ethnic issues. The Guth algorithm is straightforward to code, portable, and provides reliable results. It is, however, weak when comparing short names (Snae & Diaz, 2002), for example names like "Leon" and "Noel" are considered the same.
- Levenshtein algorithm. These are strictly alphabetic techniques based on edit distance metrics first fully described by Levenshtein (1965). Edit distance is defined for strings of arbitrary length and counts differences between strings in terms of the number of character insertions and deletions needed to convert one into the other, the minimum edit distance is then the similarity.

2) Phonetic/sound based algorithms: Methods assuming that the string representation captures sound are usually termed phonetic; however, it is important to note that there may be no explicit phoneme structure present. A North East man trying to spell a name is not capturing the phoneme construction due to his accent but is spelling the name correctly but with his perception of the sounds represented by each letter, or syllable.

Thai spelling of first name is therefore at best an approximate phonetic representation. We define phonetic methods as an attempt to follow the sound structure present in the spelled ways since these can be no “correct” or “standard” spelling which is invariably accurate. There are several algorithms available that assign a value to a string based on how it sounds. For example people attempt to capture sound by writing down what they have heard and they believed the way they

wrote from listening is correct e.g., "Smith" to "Smythe" or “Chakkrit” to “Jakkid”. Example algorithms of this type are “Soundex”, “Metaphone”, “Phonex”, and “NYSIIS”.

- Soundex algorithm. The method implemented here is due to Odel and Russell (1922). Soundex is a commonly used technique and has been modified for languages other than English.
- Metaphone algorithm. This type name is taken from Binstock and Rex (1995) although many variants exist. The method implemented assumes English phonetics but works equally well for forenames and surnames.
- NYSIIS algorithm is an alphabetic algorithm, which is easy to implement and which yields canonical index code similar to Soundex. However, NYSIIS differs from Soundex in that it retains information about the position of vowels in the encoded word by converting all vowels to the letter “A” (Gill, Goldacre, Simmons, Bettley, & Griffith, 1993). The NYSIIS method returns a purely alphabetic code.
- Phonex algorithm is a combination of the two methods, Soundex and Metaphone (Lait & Randell, 1998). The method was proved to give a good overall performance when applied to names in the English language.

3) Composite algorithms are inter-variation type algorithms, which are combined methods within sound based or spelling based methods. An example algorithm of this type is Index of Similarity Group (abbreviated to I.S.G.). Composite methods may result in a probability value, e.g. I.S.G. uses the probability value to identify and measure whether or not names are the same or similar in quantity.

- I.S.G. algorithm. This algorithm uses techniques combining alphabetic and phonetic approaches. The similarity comparison is based on the Guth method. The method implemented is due to Bouchard and Pouyez (1980).

4) Hybrid algorithms combine phonetic and spelling based approaches, which uses a similarity measure as probability. Examples of algorithms of these types are the LIG algorithms.

- LIG algorithms (e.g. LIG1, LIG2, and LIG3) are hybrid algorithms which combine phonetic and spelling based approaches using similarity measure as probability which described by Snae (Snae and Diaz, 2002). The algorithms are a combination of three name matching methods: Levenshtein, Index of Similarity Group (called ISG), and Guth. The LIG algorithms have the best performance in term of producing most accurate true matches, overcoming name variations, and increasing the hit rate.

Phonetic/Sound Based Algorithms

Typical phonetic algorithms basically work by suppressing the vowel information (because it is unreliable) and giving the same code to letters or groups of letters that sound the same (e.g. "PH" sounds like "F", so they are given the same code).

The main issue is relating together words that sound the same but are spelt differently; such words should have the same sound code. In the search for the most frequently used spelling of a particular name, the user would type that word and the program calculates its sound code, searches the text(s) for all words with the same code, and presents the user with the name with the greatest frequency.

There are several algorithms available that assign a value to a string based on how it sounds. However, in this reaearch two examples of algorithms are explained because we implemented a new algorithm based on following algorithms.

Soundex Algorithm

The Russell Soundex Code algorithm is a phonetic coding algorithm where names are coded in such a way that variants of the name should all receive the same code. The Soundex algorithm converts each name to a four-character code that can be used to identify equivalent names. The rules for coding a name are:

1. retain the first letter of input name and serves as the prefix letter
2. if A, E, I, O, U, Y, H and W are not initial ignored entirely
3. while (While Loop) input name is not converted to output Soundex code and it is less than 4 characters, transcode other letters to Soundex code as follows:

Letter:	letter is coded to:
B, P, F and V	1
C, G, J, K, Q, S, X and Z	2
D and T	3
L	4
M and N	5
R	6

4. output Soundex code
5. if output Soundex code is less than four characters add trailing zeros, otherwise drop rightmost characters (the remaining letters are ignored) to get the format letter, digit, digit, digit

The Soundex brings together some common variants of names such as Smith, Smeath, Smeith, Smiyth, Smitte, and Smett (these names have the same Soundex code: S530).

Metaphone Algorithm

The Metaphone method is similar to Soundex in purpose and is based on commonplace rules of English pronunciation. Many common rules of English pronunciation that Soundex does not cover, are coded in Metaphone such as in the case of 'C' pronounced as 'S' or pronounced as 'K'. Metaphone seems to have improved the hit rate on the basis of similarity. It is developed for matching words that sound alike. The algorithm ignores vowels after the first letter and reduces the remaining alphabet to sixteen consonant sounds: B X S K J T F H L M N P R Q W Y.

The algorithm converts each name to a four-character code, which is structured as follows:

1. Delete non-alphanumeric characters and change all letters to capitals (note: non-alphanumeric characters may appear in such digital records as genealogies in leading or trailing position of a name)
2. If KN or GN or PM or AE or WR are initial, drop the first letter
3. If X is initial, change to S
4. If WH is initial change to W
5. For loop C convert name to four character Metaphone code:
   5.1 Ignore duplicate letters (duplicate letters are not added to code)
   5.2 If A, E, I, O, U, and Y are initial retain them as code, otherwise delete them
   Transcode letters into sixteen consonant sounds: B X S K J T F H L M N P R Q W Y as follows:

Letter	**Coded to**	Conditions
B	B	unless at the end of a word after “m” as in “dumb”
C	X	(‘sh’ sound) if “-cia-” or “-ch-”
	S	Is “-ci-“, “-ce-“, or “-cy-“ silent if “-sci-“, “-sce-“, or “-scy-“
	K	otherwise, including “-sch-“
D	J	If in “-dge-“, “-dgy-“, “-dgi”
	T	Otherwise
F	F	
G		silent if in “-gh-“ and not at end or before a vowel in “-gn” or “-gned” in “-dge-“, etc., as in above rule
	J	If before “i”, or “e”, or “y”, if not double “gg”
	K	Otherwise
H		silent if after vowel and no vowel follows
	H	Otherwise
J	J	
K		silent if after “c”
	K	Otherwise
L	L	
M	M	
N	N	
P	F	if before “h”
	P	Otherwise
Q	K	
R	R	
S	X	(sh) if before “h” or in “-sio-“ or “-sia-“
	S	Otherwise
T	X	(sh) if “-tia-“ or “-tio-“
	0 (zero)	(th) if before “h” silent if in “tch-”
	T	Otherwise
V	F	
W		silent if not followed by a vowel
	W	if followed by a vowel
X	KS	
Y		silent if not followed by a vowel
	Y	if followed by a vowel
Z	S	

6. Return Metaphone Code

The Metaphone brings together some common variants of names, such as Smith, Smeath, Smeith, and Smiyth (Metaphone code of these names are SM00), whereas for Smitte and Smett it returns SMT0 as a Metaphone code.

A Novel Phonetic Name Matching Algorithm

A new composite name matching technique called MetaSound (combination between Soundex and Metaphone algorithms) has also been implemented and is used for finding name variants (spelling and phonetic variations). Since Soundex and Metaphone are phonetic coding algorithms which are designed primarily for only use with English names. Therefore, MetaSound is developed based on commonplace rules of Thai pronounciation for matching words that sound and are spelled alike. The algorithm reduces the Thai alphabet to eight consonant sounds: : /k/ or /kh// K, /b/ or /p/ B, /d/, /t/ or /th/ D, /ŋ/ NG, /n/ N, /m/ M, /j/ Y and /w/ W as follows (Snae et al., 2007):

1. retain the first letter of input name and serves as the prefix letter
2. vowels are ignored entirely
3. ignore mute pseudo-cluster combinations (Commission on Higher Education, 2006) + letters with the mute indicator, called Karan symbol (□ □)
4. while (While Loop) input name is not converted to output MetaSound code and it is less than 4 characters, transcode other letters to MetaSound code as follows:

Letter	Sound	Letter is coded to
ก ข ค ฅ ฆ	K	1
จ ฉ ช ซ ฌ ฎ ฏ ฑ ด ต ถ ท ธ ศ ส ษ	D	2
บ ป พ ฟ ผ ฝ ภ	B	3
ง	NG	4
น ล ฬ ร ณ ญ	N	5
ม	M	6
ย	Y	7
ว	W	8

5. output Soundex code to the form letter, digit, digit, digit \
6. if output Soundex code is less than four characters add trailing zeros, otherwise drop rightmost characters (the remaining letters are ignored)
7. return MetaSound Code

Figure 1 shows the use of MetaSound algorithm for name variations and matching in the proposed Thai naming system. For that we use a dictionary database of more than 10,000 Thai names which contains not only the spelling, but also the meaning and correct pronunciation to compare the name with similar names in a dictionary database and check for similarity and high rate of good name according to Thai astrology.

The program code of the MetaSound algorithm can be seen in the appendix, and the whole system uses PHP and SQL for the web-based and database subsystems.

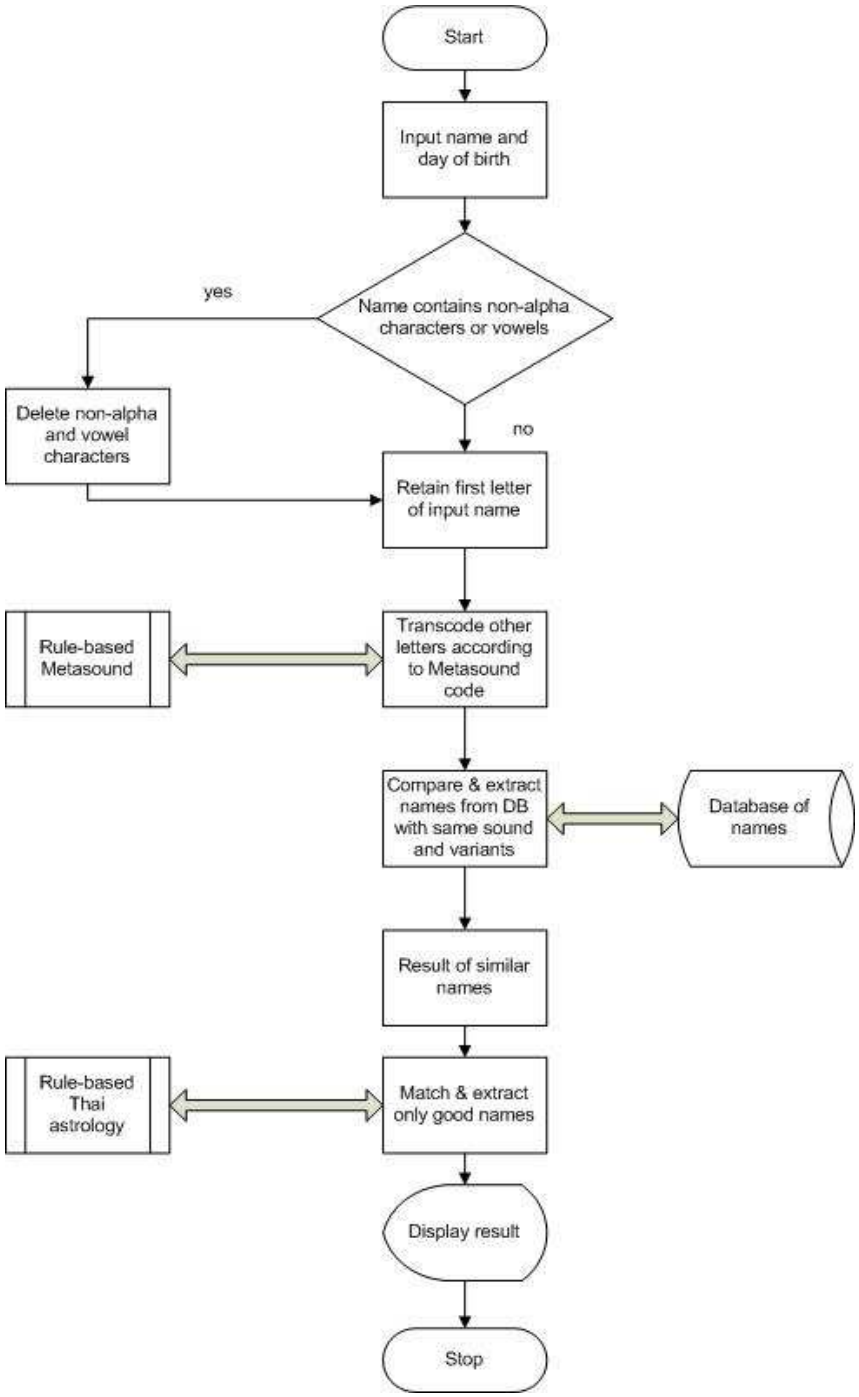


Figure 1 Flowchart of Thai Naming System using MetaSound

Results of Thai Name Checking System using Statistical Ontology

Figure 2 shows one example of the test results for the Thai name checking system using a statistical ontology. It is indicated that "กนกกาญจน์" (pronounced as Kanokkarn) has an R value of 100%.

This means that this name is a perfect name according to the Thai astrology and contains no letter from the set of misfortune letters:

ก	k / k / k	attribute Power
า	a	attribute Age
น	n / n	attribute Diligence
ญ and จ	n / (mute t)	attribute Honour

3 letters in attribute Power, then the goodness of this attribute is (3/9)*100 = 33.33 %.

2 letters in attribute Honour and Diligence, then the goodness of these attributes is (2/9)*100 = 22.22 %.

Age attribute has one letter, so the goodness value of the four attributes is (1/9)*100 = 11.11%. Therefore the statistical goodness (G) of this name (kanokkarn) is 33.33+22.22+22.22+11.11 = 89% (Og =2) and this name has no misfortune attribute (Om =1) therefore this name has R = (2+1)/3 =100%.

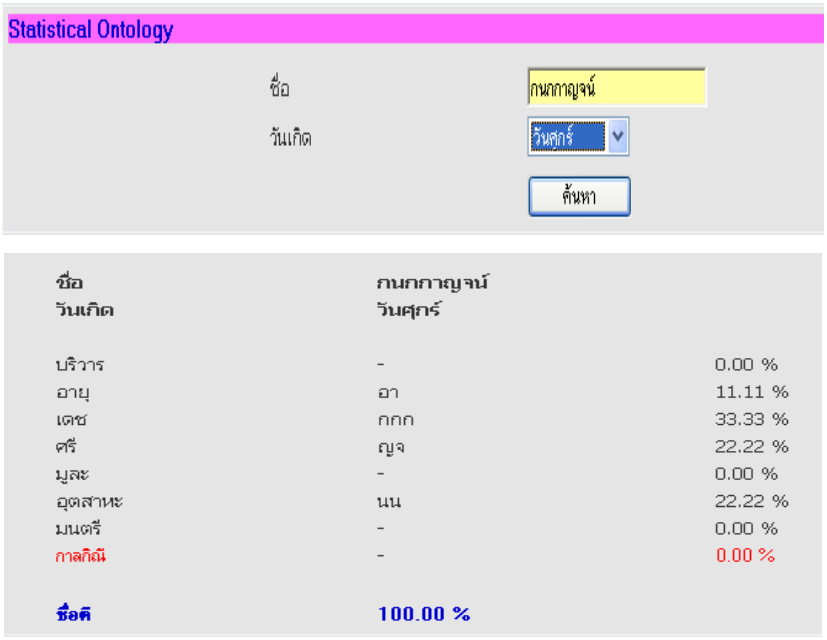
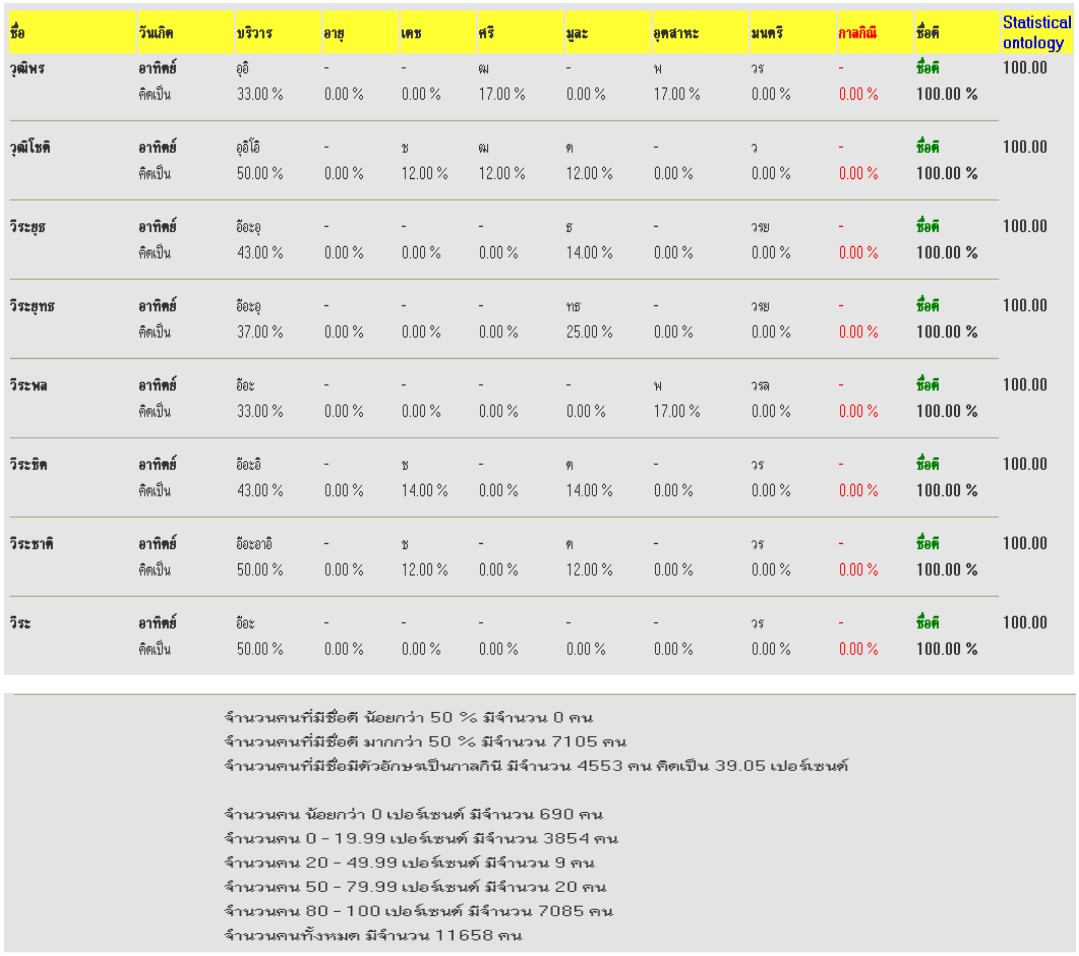


Figure 2 Results of Name Checking System for กนกกาญจน์ (Kanokkarn) who was born on Friday

Thai name checking system using statistical ontology concept has been tested with a database of 11,658 student names of Naresuan University to analyse whether students are named according to Thai astrology, and the results are shown in Figure 3.

From Figure 3 it it can be seen that 7,105 students do have good names according to Thai astrology with R greater than 50%. The percentage of students that have names without a misfortune attribute is 61.00 %. This indicates that most of the parents of the students were concerned how to name their children according to the Thai astrology.



ชื่อ	วันเกิด	บริวาร	อายุ	เดช	ศรี	มูละ	อุตสาหะ	มนตรี	กาลกิณี	ชื่อดี	Statistical ontology
วุฒิพร	อาทิตย์	อุอิ	-	-	ฒ	-	พ	วร	-	ชื่อดี	100.00
	คิดเป็น	33.00 %	0.00 %	0.00 %	17.00 %	0.00 %	17.00 %	0.00 %	0.00 %	100.00 %	
วุฒิโชติ	อาทิตย์	อุอิโอิ	-	ช	ฒ	ต	-	ว	-	ชื่อดี	100.00
	คิดเป็น	50.00 %	0.00 %	12.00 %	12.00 %	12.00 %	0.00 %	0.00 %	0.00 %	100.00 %	
วิระยุธ	อาทิตย์	อิอะอุ	-	-	-	ธ	-	วย	-	ชื่อดี	100.00
	คิดเป็น	43.00 %	0.00 %	0.00 %	0.00 %	14.00 %	0.00 %	0.00 %	0.00 %	100.00 %	
วิระยุทธ	อาทิตย์	อิอะอุ	-	-	-	ทธ	-	วย	-	ชื่อดี	100.00
	คิดเป็น	37.00 %	0.00 %	0.00 %	0.00 %	25.00 %	0.00 %	0.00 %	0.00 %	100.00 %	
วิระพล	อาทิตย์	อิอะ	-	-	-	-	พ	วล	-	ชื่อดี	100.00
	คิดเป็น	33.00 %	0.00 %	0.00 %	0.00 %	0.00 %	17.00 %	0.00 %	0.00 %	100.00 %	
วิระธิต	อาทิตย์	อิอะอิ	-	ช	-	ต	-	วร	-	ชื่อดี	100.00
	คิดเป็น	43.00 %	0.00 %	14.00 %	0.00 %	14.00 %	0.00 %	0.00 %	0.00 %	100.00 %	
วิระชาติ	อาทิตย์	อิอะอาอิ	-	ช	-	ต	-	วร	-	ชื่อดี	100.00
	คิดเป็น	50.00 %	0.00 %	12.00 %	0.00 %	12.00 %	0.00 %	0.00 %	0.00 %	100.00 %	
วิระ	อาทิตย์	อิอะ	-	-	-	-	-	วร	-	ชื่อดี	100.00
	คิดเป็น	50.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	0.00 %	100.00 %	

จำนวนคนที่มีชื่อดี น้อยกว่า 50 % มีจำนวน 0 คน
จำนวนคนที่มีชื่อดี มากกว่า 50 % มีจำนวน 7105 คน
จำนวนคนที่มีชื่อมีตัวอักษรเป็นกาลกิณี มีจำนวน 4553 คน คิดเป็น 39.05 เปอร์เซนต์

จำนวนคน น้อยกว่า 0 เปอร์เซนต์ มีจำนวน 690 คน
จำนวนคน 0 - 19.99 เปอร์เซนต์ มีจำนวน 3854 คน
จำนวนคน 20 - 49.99 เปอร์เซนต์ มีจำนวน 9 คน
จำนวนคน 50 - 79.99 เปอร์เซนต์ มีจำนวน 20 คน
จำนวนคน 80 - 100 เปอร์เซนต์ มีจำนวน 7085 คน
จำนวนคนทั้งหมด มีจำนวน 11658 คน

Figure 3. Results of name checking system using statistical ontology and clustering for 11,658 student names

Results of Thai Naming System using MetaSound

We have constructed and implemented a Thai naming system using MetaSound and Thai astrology, which offers a basic way to come to “good” Thai names according to the Thai astrology methodology. This can help users to change their names if they are not satisfied by their old names because either having a low rating or misfortune attribute.

The MetaSound application helps to find name variants that are similar to the old one that users would like to change for re-naming themselves, which is quite popular among Thais in case of misfortune in their lives.

We use an input name that is transcoded into MetaSound code using rule based MetaSound. For example, users would like to change name from บูรณะ (Burana). Then we compare the different results with our database and return the matches. The user will be able to choose from a list of resulting similar names according to their respective meaning. Users have also to select the day of birth (e.g. Thursday). The day of birth is used for comparing with rule base Thai astrology for checking good names which would not contain any misfortune letters, e.g. บัณฑอน (Bunthon) and บุณฑรา (Bunthara).

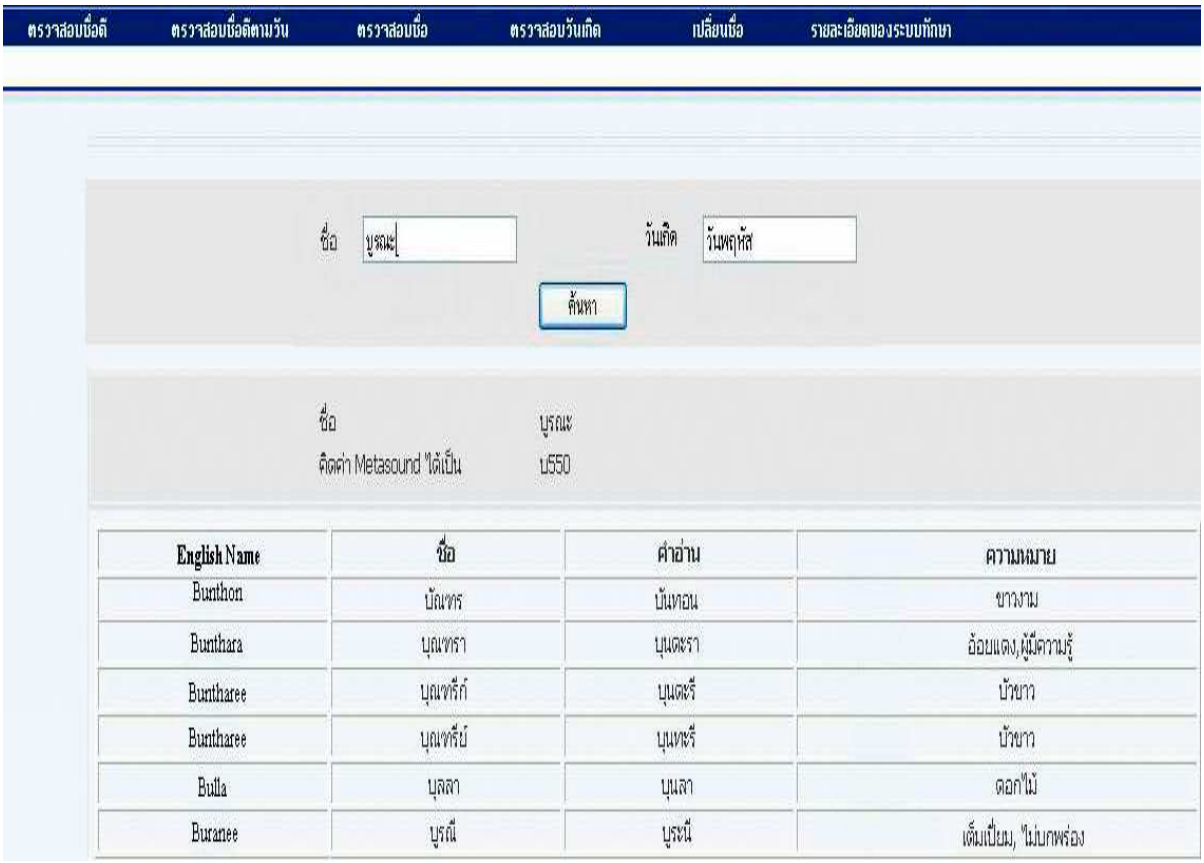


Figure 4 Results of names that are similar to a given one

Conclusion

In this research we have developed Thai astrology name checking and Thai naming systems using a novel name matching algorithm and a statistical ontology with clustering for checking and finding good names according to the Thai astrology concept. The Thai name checking system based on the value of a statistical ontology was to help defining a group of statistical values and to improve the reliability of Thai name checking. We have used Thai astrology as a naming methodology, the MetaSound algorithm for personal name matching to return the variants of names from a database with the relative probability of their similarity.

Our future intention is to implement an English naming system based on the Thai astrology naming concept for foreigners e.g., using a name transliteration system (Snae et al., 2006; Snae and Pongcharoen, 2007) to transcribe English names to the Thai writing system before checking for the goodness of the names. A prototype result of this "Romanized" system can be seen in Figure 5.

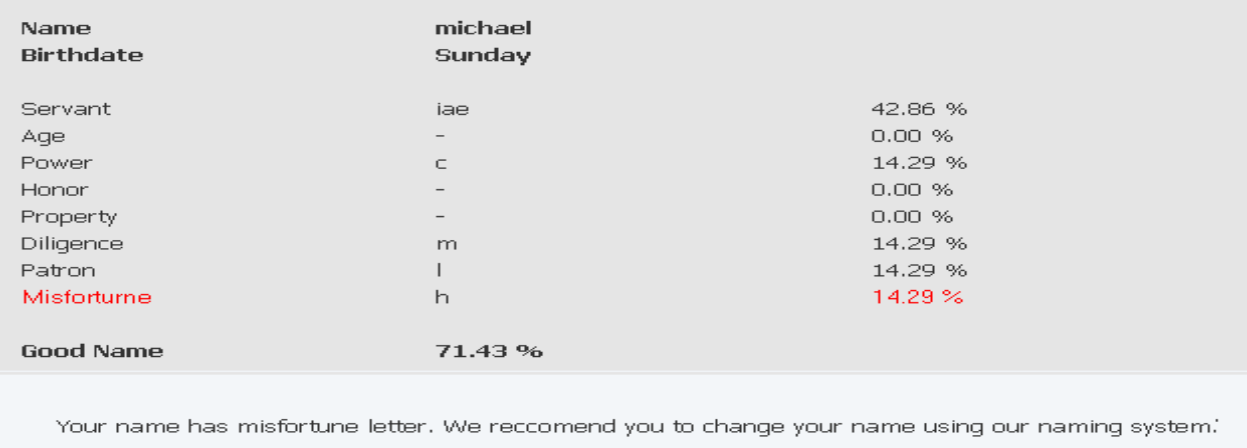


Figure 5 Output of the experimental Romanized name checking system according to Thai astrology

In addition, the Thai astrology naming system provides several choices for users: to choose the letters for each date of birth for constructing names, to combine different letters, syllables and names from parents' names into a new name, to find similar names to their names, and finally to choose names from meaning. Finally, we would use Expert System to help in name checking and naming systems e.g., if users checked their names which had low statistical values and names had misfortune attribute then the system will suggest and display good names with the day of birth that have high statistical values with no misfortune in them.

References

Binstock, A., & Rex, J. (1995). *Practical algorithms for programmers* (158-160). Addison-Wesley, Reading, Mass.

Blackburn, S. (1996). *The Oxford dictionary of philosophy*. Oxford: OUP.

Bouchard, G. & Pouyez, C. (1980). Name variations and computerised record linkage. *Historical Methods, 13*(2), 119-125.

Denk, M., Froeschl, K. A., & Grossmann, W. (2002). Statistical composites: A transformation-bond representation of statistical data. In *Proceeding 14th Conf, ACM SIGMOD*, Los Alamitos, 219-226.

Gill, L. E., Goldacre, M. J., Simmons, H. M., Bettley, G. A & Griffith, M. (1993). Computerised linkage of medical records: Methodological guidelines. *Journal of Epidemiology and Community Health, 47*, 316-319.

Gruber, T. R. (1993). A translation approach to portable ontology specification. *Knowledge Acquisition, 5*(2), 199-220.

Guth, G. J. A. (1976). Surname spellings and computerized record linkage. *Historical Methods Newsletter, 10*(1), 10-19.

Hert, C. A., & Haas, S. (2003). Support end-users of statistical information: The role of statistical metadata in the statistical knowledge network. *International Conference Proceeding Series.130*.

Lait, A. J., & Randell, B. (1998) An assessment of name matching algorithm. *Society of Indexers Genealogical Group, Newsletter Contents, SIGGNL*, Issues 17.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR, 163*, 845-848 (trans. *Soviet Physics Doklady, 10*, 707-710).

Lorchirachoonkul, V. (1982). A Thai soundex system. *Information Processing and Management, 18*(5), 243-55.

Marchionini, G., Haas, S., Plaisant, C., Shneiderman, B. & Hert, C. (2003). Towards a statistical knowledge network. *Proceedings of the National Conference on Digital Government Research*.

Odell, K. M. & Russell, R. C. (1922). *Soundex phonetic comparison system* [cf. U.S. Patents 1261167 (1918), 1435663 (1922)].

Pasquier, C., Girardot, F., Jevardat de Fombelle, K., & Christen, R. (2004). THEA : Ontology-driven analysis of microarray data. *Bioinformatics, 20*(16), 2636-2643.

Snae, C. (2006). Constructing a rule based naming system for Thai names using the concept of ontogies. *11thConference of the European Chapter of the Association for Computational Linguistics* (EACL 2006), Learning Structured Information in Natural Language Applications, Italy: Trento.

Snae, C. (2007). A comparison and analysis of name matching algorithms. *International Journal of Applied Science, Engineering and Technology, 4*(1), 252-257.

Snae, C., & Brueckner, M. (2006a). Concept and rule based naming system. *Journal of Issues in Informing Science and Information Technology, 3*, 619-634. Retrieved from http://informingscience.org/proceedings/InSITE2006/IISITSnae179.pdf

Snae, C., & Brueckner, M. (2006b). Hybrid name matching methods for rule based Thai naming system. *Naresuan University Science Journal, 2*(2), 139-150.

Snae, C., & Diaz, B. M. (2002). An interface for mining genealogical nominal data using the concept of linkage and a hybrid name matching algorithm. *Journal of 3D-Forum Society, 16*(1), 142-147.

Snae, C., & Namahoot, K. (2007). Web-based Thai name checking system using Thai astrology naming and clustering techniques. *Naresuan University Science Journal*. Retrieved from http://www.sci.nu.ac.th/journal/

Snae, C., Namahoot, K., & Brueckner, M. (2007). MetaSound: A new phonetic based name matching algorithm for Thai naming system. *International Conference on Engineering, Applied Science and Technology (ICEAST 2007)*, 21-23 November, Bangkok, Thailand.

Snae, C., & Pongcharoen, P. (2007). Automatic rule-based expert system for English to Thai transcription (RESETT) *The IASTED International Conference on Advance in Computer Science and Technology (ACST2007)*, Phuket, Thailand, 2 – 4 April.

Snae, C., Singhadech, N., Emapana, B., & Brueckner, M. (2006) Interactive transliteration tools for explanation level language system (IT-TELLS). *Proceedings of the ITC-CSCC 2006: The 21st International Technical Conference on Circuits/Systems, Computers and Communications*, Chiang Mai, Thailand.

Appendix: Program Code for the MetaSound algorithm

```
function MetaSound($name){
$name1=$name;
$count=strlen($name1);

for($j=0;$j<=$count;$j++)
{

if($name1[$j]=='ก'||$name1[$j]=='ข'||$name1[$j]=='ฃ'||$name1[$j]=='ค'||$name1[$j]=='ฅ'||$name1[
$j]=='ฆ' ||$name1[$j]=='ง'||$name1[$j]=='จ'
||$name1[$j]=='ฉ'||$name1[$j]=='ช'||$name1[$j]=='ซ'||$name1[$j]=='ฌ'||$name1[$j]=='ญ'||$name1[
$j]=='ฎ'
||$name1[$j]=='ฏ'||$name1[$j]=='ฐ'||$name1[$j]=='ฑ'||$name1[$j]=='ฒ'||$name1[$j]=='ณ'||$name1[
$j]=='ด'
||$name1[$j]=='ต'||$name1[$j]=='ถ'||$name1[$j]=='ท'||$name1[$j]=='ธ'||$name1[$j]=='น'||$name1[$
```

```
j]=='บ'
||$name1[$j]=='ป'||$name1[$j]=='ผ'||$name1[$j]=='ฝ'||$name1[$j]=='พ'||$name1[$j]=='ฟ'||$name1[$
j]=='ภ'
||$name1[$j]=='ม'||$name1[$j]=='ย'||$name1[$j]=='ร'||$name1[$j]=='ล'||$name1[$j]=='ว'||$name1[$j
]=='ศ'||$name1[$j]=='ฤ'||$name1[$j]=='ฦ'
||$name1[$j]=='ษ'||$name1[$j]=='ส'||$name1[$j]=='ห'||$name1[$j]=='ฬ'||$name1[$j]=='อ'||$name1[
$j]=='ฮ'||$name1[$j]=='ั')
                        $word[$j]=$name1[$j];
                }
                for($j=0;$j<=$count;$j++)
{
                if($word[$j]=='ั'){

                $word[$j-1]=" ";
                $word[$j]=" ";
                }
}

                for ($i=1;$i<=$count;$i++)
                        {
if($word[$i]=='ก'||$word[$i]=='ข'||$word[$i]=='ฃ'||$word[$i]=='ค'||$word[$i]=='ฆ'||$word[$i]=='ฅ')
                                        {
                                        $name1[$i]='1';
                                        }
elseif($word[$i]=='จ'||$word[$i]=='ฉ'||$word[$i]=='ช'||$word[$i]=='ฌ'||$word[$i]=='ซ'||$word[$i]=
='ฐ'
                                         ||$word[$i]=='ท'||$word[$i]=='ฒ'||$word[$i]=='ด'
||$word[$i]=='ฎ'||$word[$i]=='ต'||$word[$i]=='ส'||$word[$i]=='ศ'||$word[$i]=='ย')
                                         {
                                        $name1[$i]='2';
                                         }
elseif($word[$i]=='ฟ'||$word[$i]=='ฝ'||$word[$i]=='พ'||$word[$i]=='ผ'||$word[$i]=='ภ'||$word[$i]=
='บ'||$word[$i]=='ป')
                                        {
                                        $name1[$i]='3';
                                        }
                                        elseif($word[$i]=='ง')
                                        {
                                        $name1[$i]='4';
                                        }
elseif($word[$i]=='ล'||$word[$i]=='พ'||$word[$i]=='ร'||$word[$i]=='น'||$word[$i]=='ฌ'||$word[$i]=
='ฦ'||$word[$i]=='ญ')
                                        {
                                        $name1[$i]='5';
                                        }
```

```
                elseif($word[$i]=='ม')
                {
                $name1[$i]='6';
                }
                 elseif($word[$i]=='ย')
                {
                $name1[$i]='7';
                }
elseif($word[$i]=='ว')
                {
                $name1[$i]='8';
                }
               else
                {
                $name1[$i]='0';
                }
```

Biographies

Chakkrit Snae is currently a lecturer at Department of Computer Science and Information Technology, Naresuan University, Thailand. He is also a head of the research group KIND-HEART (Knowledge-based Intelligent systems using Natural language processing, Data mining, Heterogeneous ontologies and Expert system Application, Research and Technology).

He received a Ph.D. in Computer Science from University of Liverpool, England, M.Sc. in Computer Science from University of Newcastle upon Tyne, England, and B.Sc. in Mathematics from Naresuan University, Thailand. His research interests include Web Based Technologies, System Applications, Semantic Web, Ontologies, Machine Learning, Data/Web Mining, Software Engineering, Intelligent and Expert Systems.

Michael Brueckner is currently lecturer at the Faculty of Science, Naresuan University, Phitsanulok, Thailand, and deputy head of the research group KIND-HEART (Knowledge-based Intelligent systems using Natural language processing, Data mining, Heterogeneous ontologies and Expert system Application, Research and Technology).

He earned a diploma in physics (Dipl. Phys.) from the Technical University Munich, Germany, and worked on simulation software for physical processes, Computer-Aided Design, project management and software quality assurance. He is involved in the field of knowledge and information management since more than two decades. His current research interests are Semantic Web technologies, GIS, ontologies, natural language processing, and intelligent systems.