Bożena Śmiałkowska, Marcin Gibert

The base of knowledge in the Internet domain name rating system by using extraction and inflectional classification of words from domain names

Ekonomiczne Problemy Usług nr 106, 345-358

2013

ZESZYTY NAUKOWE UNIWERSYTETU SZCZECIŃSKIEGO

NR 781 EKONOMICZNE PROBLEMY USŁUG NR 106 2013

BOŻENA ŚMIAŁKOWSKA
MARCIN GIBERT
West Pomeranian University of Technology

THE BASE OF KNOWLEDGE IN THE INTERNET DOMAIN NAME RATING SYSTEM BY USING EXTRACTION AND INFLECTIONAL CLASSIFICATION OF WORDS FROM DOMAIN NAMES

Introduction

The so-called mentally method is the currently used method for the rating of Internet domain names by specialists. However, this method proves to be risky, as shown by having difficulty analysing large amounts of data, a limited and subjective selection of criteria for the rating of domain names and finally a changeable variability in time of the domains, as mentioned in literature[1].

The most important variable influencing the value of an Internet domain name is the kind of word form, the so-called keywords[2]. However, using the Rough Set Theory (RST)[3] and/or Fuzzy Set Theory (FST)[4], there is no automatic analysis of words from domain names. This increases the risks of rating.

[1] D. Kesmodel, *The Domain Game: How People Get Rich from Internet Domain Names,* Xlibris Corporation, Philadelphia 2008, pp. 79, 185.

[2] M. Gibert, B. Śmiałkowska, *Wycena domen internetowych z wykorzystaniem teorii zbiorów przybliżonych*, Metody Informatyki Stosowanej 2010, No. 4 (25), pp. 11.

[3] L.A. Zadeh, *Granular Computing and Rough Set Theory*, Lecture Notes in Computer Science 2007, Vol. 4585, pp. 1-4.

[4] A. Piegat, *Fuzzy modeling and control*, Physica, Heidelberg 2001, pp. 2.

Therefore the expanded system of the inflectional classification of words is used in this article. This system is based upon the tree of inflectional classification of words in the Polish language. Therefore the three-layer mechanism, representing the inflectional dictionary, which recognizes one- and more-segment words, has been described and the algorithm for extracting words from domain names which uses this system of inflectional classification has been developed. Finally the linked RST method with developed algorithm has been presented by authors as an example of rating an Internet domain name.

1. Characterization of the grammatical category tree of the Polish language

A very important element of the rating by using a method of data-exploration is the selection of correct criteria. Concerning Internet domain names there are a lot of criteria which have a strong influence on the rating. These are, among others, the age of a domain, the extension, kind of domain, PageRank, mentioned in literature[5].

One of the most important variables influencing the value of Internet domain names is the kind of word form, the so-called keywords. A correct analysis of domain names by using extraction and inflectional classification of words in these domain names allows to obtain additional criteria which influences the rating of a domain name[6]. The inflectional variant determines the classification of words in the Polish language. By transforming a basic word different inflectional word forms are derived. This transformation adds inflectional extensions to an inflectional subject. The meaning of the word is represented by the inflectional subject. On the other hand the particular value of a grammatical category of words is being determined by the inflectional extension. This grammatical category of words depends on two main processes of inflectional variation. The first is declination – variation by case – and the second one is conjugation – variation by persons and times. The joint part of declination

[5] B. Anderson, *Making money from domain names - the domain flipper's bible*, Lion Pride Publishing co. LLC 2012, pp. 24.

[6] M. Gibert, B. Śmiałkowska, *Method for making decisions on investing on the Internet domain market with use of the fuzzy sets theory*, Studia i Materiały Polskiego Stowarzyszenia Zarządzania Wiedzą 2011, No. 57, pp. 373.

and conjugation is a grammatical category of the number, singular or plural. In the declination also exists the grammatical category of gender. Concerning conjugation are distinguished the category of time and the categories of mood, side and gender. In the Polish language are distinguished parts of speech, based upon criteria of syntax and inflectional word-classes. These are: A – nouns, B – verbs, C – adjectives, D – numerals, E – pronouns, F – adverbs, G – invariables. Every word-class is divided into some subgroups based upon the inflectional extension of each word-class, e.g. nouns are divided into three subgroups, according to the gender: masculine, feminine, neuter. Subsequently they are further divided into mood and number. The division of words into word-classes is the top of the grammatical category tree, presented in literature[7]. In this tree every pattern of variation of the Polish language is ordered in a logic way. Every node of the tree is described by a sequence of letters. By descending the tree a letter combination is gathered, which identifies a grammatical category and inflectional extension of words. The word-classes representing the highest level in the tree, are followed by the gender of grammatical variation of the words. For a noun the levels of the tree are respectively: A – masculine, B – animate, C – inanimate, D – feminine, E – neuter, F – Personal, G – impersonal, H – indeclinable. The label of words is complete after determining all these levels of the tree. The example of the word "kot" (cat) results in the label ABABAB, which means: A – Noun, B – Animate, A – nominative of singular – extension is missing, B – nominative of plural - extension "y" (s), A – genitive of singular - extension "a", B – dative of singular - extension "u".

2. Description of the Inflectional Dictionary of the Polish Language

The tree of grammatical categories is the base of knowledge used by the Inflectional Dictionary of the Polish Language, which was developed and described in literature[8, 9]. The Inflectional Dictionary of the Polish

[7] *Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu*, ed. W. Lubaszewski, Wydawnictwa AGH, Kraków 2009, p. 44.

[8] *Ibidem*, p. 107.

[9] W. Lubaszewski, H. Wróbel, M. Gajęcki, B. Moskal, A. Orzechowska, P. Pietras, P. Pisarek, T. Rokicka, *Słownik fleksyjny języka polskiego*, Wydawnictwo Prawnicze LexisNexis, Warszawa 2001.

Language is presented by a three-level mechanism, which is used to recognize the one- and more-segment and potential many-segment words. The first level of the dictionary is a complete set of inflectional forms, inflectional descriptions and identifiers of variation patterns for every word included in the inflectional dictionary. The pattern is a scheme of variation of a specific group of words. Formal notation of patterns of variation is presented in literature[10]. The pattern has been described by the sets O, F, T.

$$W = \{O, F, T\},$$

where:
W – variation pattern,
O – description of inflectional forms,
F – aggregation of inflectional extensions,
T – inflectional transformation.

The description of inflectional forms of pattern W, named O, is the set of combinations of grammatical values for a specific groups of words.

$$O = \{o1, o2, ..., on\},$$

where:
o1, o2, ..., on – combinations of grammatical values.

Combinations of grammatical values, e.g. for a noun, are nominative of singular number, genitive of singular number, dative of singular number etc. Every aggregation of inflectional forms is corresponding with a set of inflectional extensions, named F.

$$F = \{f_1, f_2, \ldots, f_n\},$$

where:
f – inflectional extensions.

The inflectional transformations T, called rules of local grammars, are the full description of a pattern.

[10] W. Lubaszewski, *Słowniki komputerowe*..., pp. 39.

$$T = \{(t_1^1, t_2^1),(t_1^2, t_2^2),...(t_1^n, t_2^n)\},$$

where:
t – free form of characters string.

The rules of local grammars are ordered into pairs of character strings e.g. (r, rz), (t, ci), (s, si). Every pair describes one of the possible transformations of the inflectional subject in a pattern, which follows the context of the elements O and F. The local grammars work only in a relevant pattern. The transformation of the inflectional subject has been presented in the example of the Polish word profesor. For the word profesor with label AAAAAA the variation pattern with the inflectional extensions and the rule of subject change (r, rz) has been described.

$$W = \{\{NS, GS, DS, AS, AbS, LS, VS, NP, GD, DP, AP, AbP, LP, VP\},$$
$$\{0, a, owi, a, em, e, e, owie, ów, om, ów, ami, ach, owie\},$$
$$\{(r, rz)\}\},$$

where:
N, G, D, A, Ab, L, V – cases: Nominative, Genitive, Dative, Accusative, Ablative, Locative, Vocative,
S – singular, P – plural.

The letter "r" in the inflectional subject profesor is changed into "rz" what is derived from the local grammar (r, rz). This happens after the nominative and vocative of the singular word.

The next layer of representation of the Inflectional Dictionary of the Polish Language is a characteristic description of many-segment words. The many-segment words are developed from two or more one-segment words. The meaning of many-segment words is determined by the composing segments. An example of a many-segment word is konik morski ("seahorse"). The generation of a dictionary of many-segment words is possible because of the formal description of every dictionary item. The description of many-segment words includes:

- keyword form – describes each segment of the keyword, the spelling and punctuation,

- structural description – describes the position and order of segments, which comes first etc.,
- inflectional description – describes the inflection of the segments,
- grammatical description – characterizes word-classes,
- description of all variable forms – describes variants of words, e.g. own names like PKS are written in capital letters.

The final layer of representation of the dictionary is the regular formula of the description of potential many-segment words. Examples of potential many-segment words are: dwadzieścia cztery (twenty-four), zielono-fioletowo-czarny (green-violet-black) etc. Because of the unlimited quantity of potential many-segment words it is not possible to enclose it into a lexicon of a dictionary. This type of words in a dictionary is represented by a mechanism operating on one-segment words. This mechanism uses a relational model of forms of potential many-segment words[11].

The other type of an inflectional dictionary, but with less precision is a dictionary based on the lexicon of all possible inflectional forms of words. Because of the expanded base of this dictionary the identification of one- and many-segment words is possible by comparison key words from a domain name with words from the lexicon. However this type of dictionaries does not include the potential many-segment words. The next limit is simple grammatical characteristics rely on the distinction of speech parts. An example of this type of dictionary with a shared base of words is the Dictionary of Polish Language – SJP.PL[12]. Because of the limits this type of dictionary is not used in the next part of the article.

3. Extraction and inflectional classification of words from domain names

The above description of the functionality of the Inflectional Dictionary of the Polish Language has been realized by Library CLP, which is written in computer language C and described in literature[13]. This library has

[11] W. Lubaszewski, *Słowniki komputerowe*..., pp. 101.

[12] www.sjp.pl (28.06.2013).

[13] W. Lubaszewski, *Słowniki komputerowe*..., pp. 107.

a layer structure. The first is the inflectional layer, which is made up by the base of primary words. The second is the layer of morphological relations. The base of primary words has been generated from the Inflectional Dictionary of the Polish Language and contents about 150,000 words. Besides, the base of primary words includes composing forms of many-segment words and abbreviations. For the automatic extraction and inflectional classification of words the algorithm which uses the Library CLP has been developed. This algorithm is presented in Figure 1.

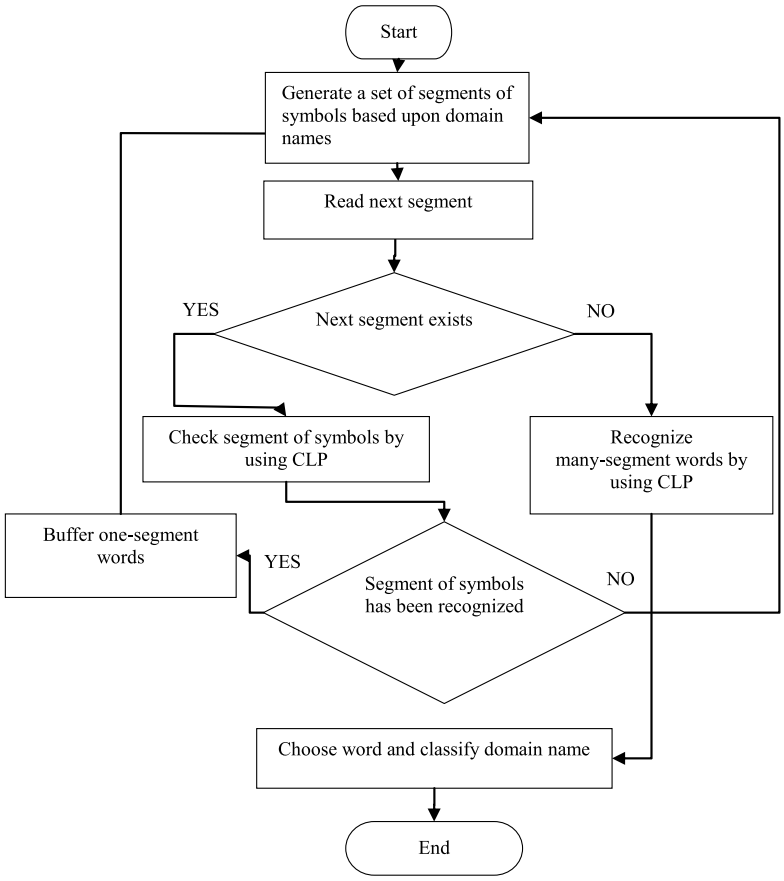

Fig. 1. The algorithm of extraction and inflectional classification of words from domain names

Source: own.

The algorithm generates every possible segment made from adjacent symbols of domain names. The specification of domain names without the diacritical characters forces the generating of alternative segments with Polish diacritical characters. The segments are candidates for identification of one-segment words. Every segment is being checked by using Library CLP. In case of recognizing a word, this word is buffered in an array of one-segment words. After the verification of all possible candidates one-segment words are checked for combining to many-segment words. In case of a positive verification, many-segment words are put into an array. In the next step is chosen from the array of many-segment words the word with the highest label. Finally, based upon the structure of names, the domain has been classified into a correct group.

4. The rating of Internet domain names by using the Rough Set Theory based upon criteria obtained from the process of extraction and inflectional classification of words

For the rating of domain names by using the RST has been used the test base from the website www.cenydomen.pl, which includes the archive with selling of Polish Internet domains. In order to make an easier calculation only a set of 15 sold domain names are used from November to December 2011. These are domains with the extension . pl and without diacritical characters. In this article have been chosen only four criteria because of limited space. These criteria are conditional attributes:

1. Length of the domain name – number of symbols
2. Continuity of registration – number of months counted from registration date in base WHOIS.
3. Type of domain name – word-class groups in accordance to extraction and inflectional classification of words from domain names.
4. Popularity – how many results are presented by Google for the keywords of a particular domain name.

The arbitrary attribute in this example is the price of a domain name. Based upon analisies of examples of sold domain names from the test base the conditional attributes *length of domain name (q1), continuity of registra-*

tion (q2), type of domain name (q3), popularity of words (q4) are represented in table 1 by the values 1, 2 and 3.

Table 1

Values of the conditional attributes

Value	**Linguistic value**	**Length of domain name (symbols)**	**Continuity of registration (months)**	**Type of domain name (segments)**	**Popularity of words (results in Google)**
1	short	fewer than 8	fewer than 3	One-segment word	less than 135000 results
2	medium	from 8 to 12	from 3 to 30	many-segment word	from 135000 to 468499 results
3	long	more than 12	more than 30	other combination	more than 468499 results

Source: own.

The arbitrary attribute *price*, d, is represented in table 2 by the values 1, 2 and 3.

Table 2

Values of the price of domain names

Value	**Linguistic value**	**Price**
1	low	Lower than 612.5 zł
2	medium	from 612.5 zł to1799 zł
3	high	More than 1799 zł

Source: own.

The test base from the archive of sales domain names[14], used in the calculations is presented in Table 3. The combination of the value numbers of the four conditional attributes results in E_i where i = 1 – 81. E_i is an elementary conditional set[15]. Based upon the conditional set E_i and arbitrary set X_i the total number of positive rules has been calculated.

[14] www.cenydomen.pl (28.06.2013).

[15] M. Inuiguchi, S. Hirano, S. Tsumoto, *Rough Set Theory and Granular Computing*, Springer, Berlin 2003, pp. 194.

Table 3

Used domain names and the values of four conditional attributes

No.	Domain name	Conditional attributes					d	
		q_1	q_2	q_3	q_4	E_i		X_i
1	demotywatory.pl	3	1	3	1	E_1	1	
2	kredytgotowkowy24.pl	3	2	3	2	E_2	1	
3	mallegro.pl	2	1	3	1	E_3	1	X_1
4	dobreubezpieczenia.pl	3	3	2	1	E_4	1	
5	krs24.pl	1	1	3	1	E_5	1	
6	agencjamultimedialna.pl	3	3	2	1	E_4	2	
7	bekon.pl	1	2	1	3	E_6	2	
8	superhity.pl	2	2	2	2	E_7	2	X_2
9	centrumkariery.pl	3	1	2	2	E_8	2	
10	elektrowniawiatrowa.pl	3	2	2	2	E_9	2	
11	pielucha.pl	2	3	1	2	E_{10}	3	
12	tramwaj.pl	1	1	1	3	E_{11}	3	
13	bombonierka.pl	2	3	1	3	E_{12}	3	X_3
14	bobaski.pl	1	2	1	3	E_6	3	
15	procesory.pl	2	3	1	3	E_{12}	3	

Source: own

Based upon the conditional set and arbitrary concepts the value $\gamma(F)$ is determined by the number of the positive examples divided into the total number of examples[16].

$$\gamma(F) = \frac{\text{card}(\text{Pos}(F))}{\text{card}(U)},$$

where:
$\gamma(F)$ – quality of approximation of arbitrary concepts F,
card(Pos(F)) – the number of positive examples,
card(U) – the total number of examples.

The next target is to determine the accuracy $\beta(F)$ of the approximation of arbitrary concepts F.

[16] X. Yang, J. Yang, *Incomplete Information System and Rough Set Theory: Models and Attribute Reductions*, Springer, Berlin 2012, pp. 135.

$$\beta(F) = \frac{\text{card}(\text{Pos}(F))}{\sum \text{card}(GP(X_i))},$$

where:
β(F) – accuracy of approximation of arbitrary concepts F,
card(Pos(F)) – the number of positive examples,
card(GP(Xi)) – quantity of upper approximation of concepts Xi.

The value β(F) is calculated by the division the total number of positive examples into the sum of the total number of examples from the elementary conditional set Ei for arbitrary set X1-3.

In the next step the whole set of attributes (q_1, q_2, q_3, q_4) is tried to be reduced by successively one attribute. Therefore the significance of the omitted attributes is calculated in table 4. The results of all mentioned calculations are also presented in table 4.

Table 4

Significance of deleted attributes

q_1	q_2	q_3	q_4	**Significance of deleted attribute (q_u)**	**Quality of concepts (γ(F))**	**Accuracy of concepts (β (F))**
q_1	q_2	q_3	q_4	-	0,733	0,579
q_1	q_2	q_3	**q_u**	0	0,733	0,579
q_1	q_2	**q_u**	q_4	0	0,733	0,579
q_u	q_2	q_3	q_4	0	0,733	0,579
q_1	**q_u**	q_3	q_4	0,091	0,666	0,5
q_1	q_2	**q_u**		0,273	0,533	0,2
q_1	q_2		**q_u**	0,273	0,533	0,2
	q_2	**q_u**	q_4	0,273	0,533	0,364
q_u	q_2		q_4	0,273	0,533	0,364
q_u	q_2	q_3		0	0,733	0,579
	q_2	q_3	**q_u**	0	0,733	0,579
	q_2	**q_u**		1	0	0

Source: own.

Based upon the calculation, shown in Table 4, two attributes, q2 and q3, which prove to be absolute reductors of the full set of conditional attributes, have been developed. The example shows that in the best case 73% of examples generate positive rules (quality), and the grade of understanding the decision is on the level of 58% (quantity). The maximum quantity of rating, based upon the use of the test base, can be obtained by using two attributes, q2 and q3, respectively *continuity of registration* and *type of the domain name*. The latter has been obtained by extraction and inflectional classification of words from a domain name. Subsequently, based upon attributes q2 and q3, unique rules have been defined. These rules are used to obtain a current value of arbitrary attribute d (price of domain). The example of rating of the domain name *przychodnielekarskie.pl* is according to the rule:

$$(q_2 = 1 \,\&\, q3 = 1) \rightarrow d = 2$$

This rule shows that *continuity of registration* (q2) is short and the *type of the domain name* (q3) is a many-segment word and because of these two attributes the price is medium. After transfering the linguistic value into a real value the price of this domain is between 612.50 zł and 1,800 zł.

Conclusion

An important element of automatic rating by using methods of data exploration, like RST, is the correct selection of criteria[17]. In case of a domain name there are many criteria which have influence on the rating. One of the most important criteria is the type of word, the so-called keyword from a domain name.

The specification of the Polish language requires the use of advanced identification technics of words from domain names, based upon the tree of the inflectional variation of words. In this article is presented the mechanism of the Inflectional Dictionary of the Polish Language to generate a correct base of knowledge for an inflectional classification of words in domain name.

[17] A. Abraham, R. Falcón, R. Bello, *Rough Set Theory: A True Landmark in Data Analysis*, Springer, Berlin 2009, pp. 72.

A three-layer structure of the inflectional dictionary has been used to recognize many-segment words. The presented system has made use of the inflectional dictionary based upon Library CLP, which allows automatic extraction and inflectional classification of words in domain names. This system allows analysing additional criteria which influence the value of domain names. The use of these criteria in the process of rating domain names, in combination with methods of data exploration like RST, improves the accuracy of rating, proved by the given example in the article.

References

Abraham A., Falcón R., Bello R., *Rough Set Theory: A True Landmark in Data Analysis*, Springer, Berlin 2009.

Piegat A., *Fuzzy modeling and control*, Physica-Verlag, Heidelberg 2001.

Anderson B., *Making money from domain names – the domain flipper's bible*, Lion Pride Publishing co. LLC 2012.

Kesmodel D., *The Domain Game: How People Get Rich from Internet Domain Names,* Xlibris Corporation, Philadelphia 2008.

Zadeh L.A., *Granular Computing and Rough Set Theory*, "Lecture Notes in Computer Science" 2007, Vol. 4585.

Gibert M., Śmiałkowska B., *Method for making decisions on investing on the Internet domain market with use of the fuzzy sets theory*, Studia i Materiały Polskiego Stowarzyszenia Zarządzania Wiedzą 2011, No. 57.

Gibert M., Śmiałkowska B., *Wycena domen internetowych z wykorzystaniem teorii zbiorów przybliżonych*, "Metody Informatyki Stosowanej" 2010, No. 4 (25).

Inuiguchi M., Hirano S., Tsumoto S., *Rough Set Theory and Granular Computing*, Springer, Berlin 2003.

Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu, ed. W. Lubaszewski, Wydawnictwa AGH, Kraków 2009.

Lubaszewski W., Wróbel H., Gajęcki M., Moskal B., Orzechowska A., Pietras P., Pisarek P., Rokicka T., *Słownik fleksyjny języka polskiego*, Wydawnictwo Prawnicze LexisNexis, Warsaw 2001.

www.sjp.pl.

www.cenydomen.pl.

Yang X., Yang J., *Incomplete Information System and Rough Set Theory: Models and Attribute Reductions*, Springer, Berlin 2012.

EKSTRAKCJA I KLASYFIKACJA FLEKSYJNA WYRAZÓW W SYSTEMIE WYCENY DOMEN INTERNETOWYCH

Streszczenie

W artykule przedstawiono klasyfikację fleksyjną wyrazów z nazw domen internetowych. Przy użyciu opisanego systemu klasyfikacji istnieje możliwość wydobycia dodatkowych kryteriów, które wpływają na wycenę domen. System klasyfikacji opiera się na *Słowniku fleksyjnym języka polskiego*. Z tego względu został przybliżony trzypoziomowy mechanizm reprezentacji słownika fleksyjnego. Mechanizm ten jest wykorzystywany do identyfikacji wyrazów jedno- i wielosegmentowych w nazwach domen. Ostatecznie zarealizowano wycenę domeny internetowej przy użyciu klasyfikacji fleksyjnej i teorii zbiorów przybliżonych. Zaprezentowany przykład dowodzi skuteczności użycia klasyfikacji fleksyjnej w procesie wyceny domen internetowych.

Tłumaczenie Bożena Śmiałkowska i Marcin Gilbert