Available online at www.sciencedirect.com
ScienceDirect
Information Processing Letters 100 (2006) 76–82
Information Processing Letters
www.elsevier.com/locate/ipl

Determining the specificity of terms using inside–outside information: a necessary condition of term hierarchy mining

Pum-Mo Ryu *, Key-Sun Choi

Computer Science Division, KAIST, KORTERM/BOLA, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea

Received 15 November 2004; received in revised form 13 November 2005; accepted 30 November 2005

Available online 11 July 2006

Communicated by W.-L. Hsu

Abstract

This paper introduces new specificity measuring methods of terms using inside and outside information. Specificity of a term is the quantity of domain specific information contained in the term. Specific terms have a larger quantity of domain information than general terms. Specificity is an important necessary condition for building hierarchical relations among terms. If t_1 is a hyponym of t_2 in a domain term hierarchy, then the specificity of t_1 is greater than that of t_2. As domain specific terms are commonly compounds of the generic level term and some modifiers, inside information is important to represent the meaning of terms. Outside contextual information is also used to complement the shortcomings of inside information. We propose an information theoretic method to measure the quantity of terms. Experiments showed promising results with a precision of 73.9% when applied to terms in the MeSH thesaurus.

Keywords: Information retrieval; Term specificity; Term hierarchy; Information theory; Inside information; Outside information

1. Introduction

Terms are linguistic realizations of domain specific concepts and term management is a core part of domain knowledge management [7]. In this paper, we introduce a new term specificity measuring method using inside and outside information together. *Specificity* is the measure of information quantity that is contained in each term within a given domain. Because term specificity is the ability of a term to describe topics precisely, it has mainly been discussed in information retrieval researches in terms of selection of accurate index terms [1,8]. Term specificity can also be applied in the task of term hierarchy mining. Because specific terms cover a narrow range in conceptual space and tend to locate at deep levels in a term hierarchy, term specificity is a necessary condition for IS-A relations among terms in a domain (say D). That is, if a term t_1 is an ancestor of another term t_2 in a hierarchy system, H_D, derived from the domain D, then the specificity of t_1 is lower than that of t_2 in D. Based on this condition, it is highly probable that t_1 is an ancestor of t_2 in H_D, when t_1 and t_2 are semantically similar enough and the specificity of t_1 is lower than that of t_2 in D as in Fig 1. However, the specificity is not a sufficient enough condition for IS-A relations, because, for example, t_1 is not similar to t_3 on the semantic level, and t_1 is not an ancestor of t_3 even

* Corresponding author.
E-mail address: pmr@world.kaist.ac.kr (P.-M. Ryu).

0020-0190/$ – see front matter
doi:10.1016/j.ipl.2005.11.024

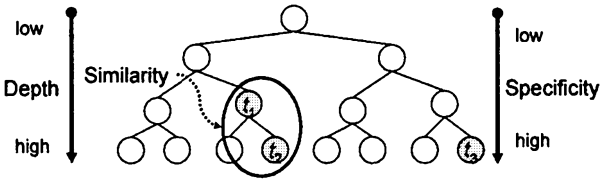

Fig. 1. Term specificity and term similarity in a domain term hierarchy H_D.

though the specificity of t_1 is lower than that of t_3 as shown in the figure.

We applied inside and outside information of terms to measure term specificity in this paper. Because many domain specific terms are multiword terms, inside information such as the characteristics of component words or the internal structure of terms is useful information to measure term specificity. Inside information has not been commonly discussed in previous researches. Those researches mainly relied on outside information such as distribution of modifiers or other statistics based on term occurrence in corpus. Caraballo [2] calculated the specificity of general nouns using the distribution of modifiers based on the assumption that specific nouns are rarely modified, while general nouns are usually modified. The purpose of this research was to aid in constructing or augmenting noun hierarchies. Aizawa [1] and Wong [8] measured term specificity based on information theoretic methods. The aim of their research was mathematical analysis of term weighting schemes commonly used in information retrieval systems. Forsyth and Rada [9] assumed that high frequency words have broad meaning, while low frequency words have narrower meanings. With this assumption they positioned the different words at the different levels of term taxonomy. They relied on term frequency to position the terms to specificity levels. Woods [10] organized term taxonomy using subsumption axioms, transitive subsumption relationship, and structural subsumption. Because the basic axioms are relied on the simple patterns, it is difficult to capture taxonomic relations that are not explicit in sentences.

Domain specific concepts have their own characteristic set. More specific concepts are created by adding other characteristics to the characteristic set of existing concepts. Let us consider two concepts: C_1 and C_2. C_1 is an existing concept and C_2 is a newly created concept by combining new characteristics to the characteristic set of C_1. In this case, C_1 is an ancestor of C_2 [6]. When domain specific concepts are embodied into terms, the terms' word-formation is classified into two categories based on the composition of component words. In the first category, new terms are created by adding modifiers to existing terms. For example "*insulin dependent diabetes mellitus*" was created by adding the modifier "*insulin dependent*" to its hypernym "*diabetes mellitus*" as in Table 1. In English, specific terms are commonly compounds of a generic level term and some modifiers [4]. In this case, inside information is a good information source to represent the characteristics. In the second category, new terms are created independently of existing terms. For example, although "*Wolfram syndrome*" is semantically related to its ancestors, it shares no common words with its ancestors. In this case, outside information is used to differentiate the characteristics of the terms. In this paper, the information of terms is quantified to be a positive real number as shown in Eq. (1).

Table 1
Subtree of the MeSH[1] tree. The node numbers represent the hierarchical structure of terms

Node number	Terms
C18.452.297	*Diabetes mellitus*
C18.452.297.267	*Insulin dependent diabetes mellitus*
C18.452.297.267.960	*Wolfram syndrome*

$$Spec(t|D) \in R^+, \tag{1}$$

where t is a term, and $Spec(t|D)$ is the specificity of t in a given domain D. As we restricted the domains into one area in this research, we simply use $Spec(t)$ instead of $Spec(t|D)$.

The remainder of this paper is organized as follows. Characteristics of inside-and-outside information are described in Section 2, while specificity measuring methods based on information theory are introduced in Section 3. Our experiments and evaluation of methods are discussed in Section 4, and finally, conclusions are drawn in Section 5.

2. Inside–outside information

2.1. Inside information

The meaning of multiword terms can be predicted from the meaning of component words. Consider a term t that consists of two words like $t = w_1 w_2$. Two words, w_1 and w_2, have their unique characteristics and the characteristics are assumed to be summed up in the making of a characteristic set of the term. We use mutual information between a term and its component words to

[1] MeSH (Medical Subject Headings) is available at http://www.nlm.nih.gov/mesh. We used the MeSH 2003 version in this research.

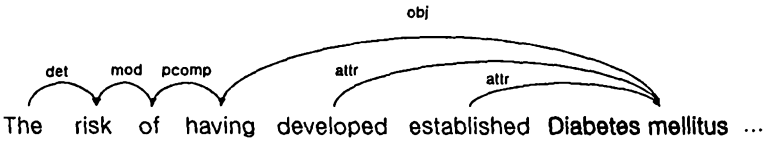

Fig. 2. The two modifiers "*developed*" and "*established*" modify the term "diabetes mellitus".

measure the information of a term. Details are discussed in Section 3.1.

Additionally, the internal modifier-head structure of terms contributes to calculating term specificity. If the structure of a multiword term is known, the specificity is calculated by adding the specificity of modifiers to the specificity of head. In this manner, the specificity of a multiword term is always larger than that of its head. However, it is very difficult to analyze the modifier-head structure of compound nouns. We use nesting relations between terms [5] to analyze the structure of terms as follows:

Definition 1. If t_1 and t_2 are terms in the same semantic category and t_1 is nested in t_2 such as $mod\, t_1$, then t_1 is the head of t_2, and *mod* is the modifier of t_1.

For example, "*diabetes mellitus*" and "*insulin dependent diabetes mellitus*" are both disease names, and the former is nested in the latter. In this case, "*diabetes mellitus*" is the head term and "*insulin dependent*" is the modifier of "*insulin dependent diabetes mellitus*". The specificity of t_2 is measured as shown in Eq. (2).

$$Spec(t_2) = Spec(t_1) + \alpha \cdot Spec(mod), \tag{2}$$

where $Spec(t_1)$ and $Spec(mod)$ are the specificity of t_1 and *mod*, respectively. They are measured using information-theoretic measure in Section 3.1. α $(0 \leqslant \alpha \leqslant 1)$ is a weighting scheme for the specificity of modifier.

Linguistic knowledge also contributes to adjusting the weight of component words. Because proper nouns, abbreviations, or other special words indicating specific classes are very informative, terms having such components tend to be located on the leaf nodes in the term taxonomy. For example, disease names like "*Wolfram syndrome*", "*HIV wasting syndrome*", and "*glycogen storage disease type I*" are found at leaf nodes in the MeSH thesaurus because they consist of very informative words: "*Wolfram*" (proper noun), "*HIV*" (abbreviation of "*Human Immune Virus*"), and "*Type I*" (special classifying word). Because most proper nouns and abbreviations start with capital letters, they are easily identified in a corpus using simple rules and statistics. Domain specific classifying words are selected manually from a domain term list. All component words are classified into two classes: high informative words and generic words based on linguistic knowledge. Varied weight values are applied to the words based on their classes in the specificity calculation process.

2.2. Outside information

There are some problems that cannot be addressed by the compositionality of multiword terms. First, although the characteristic set of "*Wolfram syndrome*" shares many common characteristics with the characteristic set of "*insulin dependent diabetes mellitus*" on a semantic level, they do not share common words on a lexical level. In this case, it is undesirable to compare the two specificity values measured using inside information alone. Second, when several words are combined in a term, there are additional semantic components that are not predicted by component words. For example, the meaning of "*diabetes mellitus*" is not predicted by the two separate words "*Wolfram*" and "*syndrome*".

Outside information can compensate these limitations. Several types of outside information are used according to object tasks such as simple collocations, the predicates of which associate target terms as their arguments, or the modifiers of target terms. Under Caraballo's assumption [2], we use probabilistic distribution of modifiers as outside information. A problem of outside information is that if a term or modifiers of the term do not occur in the corpus, the specificity cannot be measured using outside information alone. An additional problem is that because domain specific terms are rarely modified in a corpus, it is hard to collect sufficient modifiers from a given corpus. Therefore, accurate text processing such as syntactic parsing is needed.[2] Fig. 2 shows a dependency structure in which two modifiers modify the term "*diabetes mellitus*". In this case, "*developed*" is a long-dependency modifier which is hard to extract by the rightmost pronominal modifier rules used in [2].

[2] We used Conexor functional dependency parser for English (http://www.conexor.fi) to analyze the structure of sentences. Among many dependency functions defined in the Conexor parser, *attr* and *mod* functions are used to extract modifiers.

3. Specificity measuring method based on information theory

In this section, we describe information theoretic specificity measuring methods using inside and outside information. In information theory, when a low probability message occurs on a channel output, the amount of *surprise* is large, and the length of bits to represent the message becomes long. Therefore, a large quantity of information is gained by this message [3]. If we regard the term t_i found in a corpus as a message observed at a channel output, the information quantity of the event of t_i is observed, $I(x_i)$, and can be measured based on information theoretic methods. The value is assigned as the specificity of t_i, $Spec(t_i)$, as in Eq. (3). The problem of measuring the specificity of a term is then changed into the problem of calculating the information of an event of the term observed.

$$Spec(t_i) \approx I(x_i). \tag{3}$$

3.1. Inside information-based method (Method 1)

In this section, we describe a specificity measuring method using inside information, as was introduced in Section 2.1. Mutual information between component words and terms are used in this method. For a detailed description, let $T = \{t_1, \ldots, t_N\}$ be a set of terms found in a corpus, and $W = \{w_1, \ldots, w_M\}$ be a set of component words composing the terms in T. The parameters N and M are the total numbers of terms and words. Term frequency in a corpus, f_{t_i}, is assigned to all terms in T, and F_t, is the total frequencies of all terms. We assume that a term t_i consists of one or more words, and that component words in t_i differ from each other without loss of generality.[3] Therefore the frequency of a word, f_{w_j}, is equal to the total frequencies of terms which include the word. F_w is the sum of all f_{w_j} for the words in W. We define W_{t_i} as a set of words in term t_i. Fig. 3 illustrates an example corpus and the frequencies of terms and words found in the corpus.

We use x_i for the event of selecting term t_i from T, and y_j for the event of selecting word w_j from W. X and Y are random variables defined over the events $\{x_1, \ldots, x_N\}$ and $\{y_1, \ldots, y_M\}$. We also define Y_{t_i} as a set of y_j which are associated with the words in W_{t_i}. The mutual information can be used to estimate the association between terms and words. Assume a joint probability distribution $P(x_i, y_j)$ is given for $x_i \in X$ and $y_j \in Y$. Mutual information between x_i and y_j compares the probability of observing x_i and y_j together and the probabilities of observing x_i, and y_j independently as in Eq. (4).

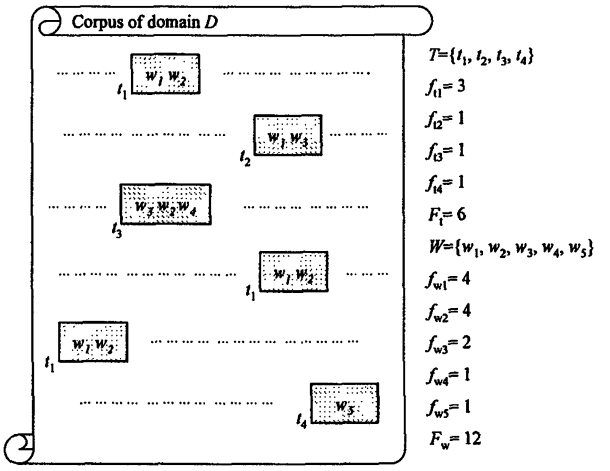

Fig. 3. An example corpus. The gray rectangles represent terms and w_js in the terms are component words.

$$I(x_i, y_j) = \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}. \tag{4}$$

The expected mutual information between x_i and Y, $I(x_i, Y)$, represents the reduction of uncertainty about x_i when Y is known. $I(x_i, Y)$ is estimated based on the frequency of terms and words in a corpus as in Eq. (5).

$$\begin{aligned} I(x_i, Y) &= \sum_{y_j \in y_{t_i}} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \\ &= \sum_{y_j \in y_{t_i}} P(x_i|y_j)P(y_j) \log \frac{P(x_i|y_j)}{P(x_i)} \\ &\approx \sum_{w_j \in W_{t_i}} \frac{f_{t_i}}{f_{w_j}} \frac{f_{w_j}}{F_w} \log \frac{f_{t_i}/f_{w_j}}{f_{t_i}/F_t} \\ &\approx \frac{f_{t_i}}{F_w} \sum_{w_j \in W_{t_i}} \log \frac{F_t}{f_{w_j}}. \end{aligned} \tag{5}$$

$I(x_i, Y)$ is used as the specificity of term t_i as in Eq. (6).

$$Spec_{in}(t_i) = \frac{f_{t_i}}{F_w} \sum_{w_j \in W_{t_i}} \beta_j \cdot \log \frac{F_t}{f_{w_j}}, \tag{6}$$

where β_j is the weighting scheme for words based on linguistic knowledge (see Section 2.1). We set $\beta_j > 1$, $\beta_j = 1$ for linguistically highly informative words and general words, respectively. f_{t_i}, f_{w_j} and β_j contribute

[3] "*itai–itai disease*" is an example of a double occurrence of the same word in a term, but such a case is rare.

to the specificity because F_t and F_w are fixed values. We can say that a term is specific when

(1) the term is composed of many component words;
(2) the frequency of each component word, f_{w_j}, is low;
(3) the frequency of the term, f_{t_i}, is high; and
(4) the term has many informative words. Here, the second and third conditions conflict with each other. If term frequency is high, word frequencies are also high. If w_j, appears only in t_i and f_{t_i} is high, specificity is maximized.

3.2. Outside information based method (Method 2)

In this section, we describe a method using outside information that was introduced in Section 2.2. Entropy of probabilistic distribution of modifiers for a term is defined as shown in Eq. (7).

$$H_{mod}(t_i) = -\sum_{1 \leqslant k \leqslant L} P(mod_k, t_i) \log P(mod_k, t_i), \quad (7)$$

where L is the number of modifiers of t_1 found in a corpus, and $P(mod_k, t_i)$ is the probability that mod_k modifies t_i. It is estimated as the relative frequency of (mod_k, t_i) in all (mod_l, t_i) $(1 \leqslant l \leqslant L)$ pairs in a corpus. The entropy is the average information quantity of all (mod_k, t_i) pairs. Specific terms have low entropy, as their modifier distributions are simple. Therefore, inversed entropy is assigned to $Spec_{out}(t_i)$ to allow specific terms a large quantity of information, as in Eq. (8).

$$Spec_{out}(t_i) \approx \max_{1 \leqslant j \leqslant N} H_{mod}(t_j) - H_{mod}(t_i), \quad (8)$$

where the first term of approximation is the maximum value among the modifier entropies for all terms.

3.3. Hybrid method (Method 3)

There are some pros and cons in the previous two methods. Method 1 reflects the characteristics of component words and Method 2 addresses a phenomenon that cannot be handled by Method 1. We introduce a hybrid method (Method 3) as delineated by Eq. (9) to combine these advantages,

$$Spec(t_i) = \frac{1}{\gamma(\frac{1}{Spec_{in}(t_i)}) + (1-\gamma)(\frac{1}{Spec_{out}(t_i)})}, \quad (9)$$

where $Spec_{in}(t_i)$, the normalized specificity value between 0 and 1, is measured by an inside information-based method using Eq. (5); and $Spec_{out}(t_i)$ the normalized specificity value between 0 and 1, is measured by an outside information-based method using Eq. (8). The value γ $(0 \leqslant \gamma \leqslant 1)$ determines the weight of the two values. If $\gamma = 0.5$, the equation is the harmonic mean of the two values. Therefore, $Spec(t_i)$ becomes large when the two values are equally large. This method is applied when both values are valid.

4. Experiments and evaluation

We applied the proposed methods to the terms in an existing thesaurus. We can say the methods are valid if the specificity values of child terms are larger than those of parent terms. The subtree of the MeSH thesaurus is selected for the experiment. "*Disease*(C)" node is the root of the subtree, and it consists of 9432 disease names. A set of journal abstracts was collected from the MEDLINE[4] database using selected disease names as search queries. Therefore, most of the abstracts are related to some of the disease names. The set consists of approximately 170 000 abstracts (20 000 000 words). The abstracts are analyzed using the Conexor parser and various statistics are extracted:

(1) the frequency of disease names;
(2) the distribution of modifiers of disease names; and
(3) the frequency of component words of disease names.

We divided parent–child relations in the subtree into two types. Relations in which parent terms are nested in child terms are categorized as Type I; and other relations are categorized as Type II. There are 1228 Type I relations and 8204 Type II relations. We can correctly measure the specificity values of terms in Type I relations, if we apply Eq. (2). We performed 5-fold experiments to find optimized values of the parameters α, β and γ. We divided the data into 5 equal parts; choose one part as the test data and the other four parts as the training data, and experimented five times with a different selection of the part for the test data. In each training run, we selected the best parameter when the precision was the highest, and applied the parameter to the test data. After five runs, we averaged the parameters.

The system was evaluated based on two criteria: coverage and precision. Coverage is the fraction of the terms that have specificity values by the given measuring method. Method 2 obtains relatively lower coverage than Method 1 because it can measure specificity when both the terms and their modifiers appear in the corpus.

[4] MEDLINE is a database of biomedical articles serviced by the National Library of Medicine, USA (http://www.nlm.nih.gov).

Table 2
The precision and coverage of Methods 1–3. In each parameter learning experiment, three types of results are displayed: (A) average results of 5 runs on training data sets using 5 locally optimized parameters, (B) average results of 5 runs on test data sets using 5 locally optimized parameters to corresponding training data, (C) average results of 5 runs on test data sets using the average of locally optimized parameters

Methods		Precision (%)			Coverage (%)
		Type I	Type II	Total	
Human subjects (average)		96.6	86.4	87.4	
Inside information method (Method 1)	MI (total data)	99.8	56.2	**61.9**	88.6
	MI + Structure (A)	99.8	66.1	70.5	94.1
	MI + Structure (B)	99.8	62.5	67.5	92.3
	MI + Structure (C) ($\alpha = 0.41$)	99.8	62.8	**67.8**	92.3
	MI + Structure + Ling. (A) ($\alpha = 0.41$)	99.7	68.9	72.9	94.1
	MI + Structure + Ling. (B) ($\alpha = 0.41$)	99.8	66.0	70.6	92.3
	MI + Structure + Ling. (C) ($\alpha = 0.41$, $\beta = 12.6$)	99.8	66.4	**70.9**	92.3
Outside information method (Method 2)		94.5	63.9	**68.0**	75.6
Hybrid method (Method 3)	(A) ($\alpha = 0.41$, $\beta = 12.6$)	99.4	72.8	75.7	75.5
	(B) ($\alpha = 0.41$, $\beta = 12.6$)	99.8	69.8	73.8	75.5
	(C) ($\alpha = 0.41$, $\beta = 12.6$, $\gamma = 0.03$)	99.8	69.9	**73.9**	75.5

Precision is the fraction of relations with correct specificity values.

$$coverage = \frac{\#\ of\ terms\ with\ specificity}{\#\ of\ all\ terms},$$
$$precision = \frac{\#\ R(p,c)\ with\ correct\ specificity}{\#\ of\ all\ R(p,c)}, \tag{10}$$

where $R(p,c)$ is a valid parent–child relation in the MeSH thesaurus, and the relation is valid when the specificity of two terms are measured by the given method. If the specificity of child term c is larger than that of parent term p, then the relation is correct.

We performed a human subject test to know the upper bound of precision. We asked 10 medical doctors to identify the parent–child relationship of given 435 term pairs. The average precisions of Types I, II and the total were 96.6, 86.4 and 87.4%, respectively. We experimented on Methods 1–3, as presented in Table 2. In Method 1, we carried out three experiments using

(1) simple mutual information-based method,
(2) structure information added method,
(3) structure information and linguistic knowledge added method, sequentially.

In the second and third experiments, we applied 5-fold parameter learning mechanism to find optimal α and β. In the second experiment, we found optimal α by repeated experiments changing α from 0 to 1 in the interval of 0.01. In the third experiment, we found optimal β by repeated experiments changing β from 0 to 100 in the interval of 1 with fixed α which was found in the second experiment. The best method in Methods 1 and 2 were combined into Method 3. In Method 3, we also applied the 5-fold learning mechanism to find optimal γ by repeated experiments changing γ from 0 to 1 in the interval of 0.01. Table 2 shows three result types in each parameter learning experiment:

(A) average results of 5 runs on training data sets using 5 locally optimized parameters in each fold,
(B) average results of 5 runs on test data sets using 5 locally optimized parameters in corresponding training data,
(C) average results of 5 runs on test data sets using the average of locally optimized parameters.

Because the results on test data are slightly lower than that on training data in the experiments, we can say that the parameters are unbiased.

In Method 1, structure information and linguistic knowledge increased the precision from 61.9 to 70.9%. This result illustrates the basic assumption that specific concepts are created by adding information to existing concepts, like the formation of multiword terms. This result also describes that the statistics extracted from the domain corpus are not sufficient to represent term specificity. Thus, other domain knowledge or linguistic knowledge is needed to represent the meaning of domain specific terms. Method 1 showed higher precision and coverage than Method 2. This result indicates that inside information is more informative than outside information. Especially coverage in Method 2 was much lower than that of Method 1. The reason is that because domain specific terms are rarely modified from

other words, we could not collect statistically sufficient modifiers from corpus. Method 3, a hybrid method of Method 1 (MI of terms and component words, structure information, linguistic knowledge) and Method 2, showed the best precision, 73.9% on test data, because the two methods interacted in a complementary manner throughout the process. However, the coverage was slightly lower than that of Method 2 since Method 3 can measure hybrid specificity when two specificities are valid. The precision of 73.9% is promising compared to the upper bound of 87.4% although still remained much to do.

5. Conclusions

We proposed new specificity measuring methods for terms based on inside and outside information using information theoretic measures. Because many domain specific terms are multiword terms, inside information contributes to measuring the information quantity of terms. Outside information such as probabilistic distribution of modifiers is also contributes to term specificity. The hybrid method, based on both inside and outside information, showed the best precision. Because specificity is a necessary condition for term hierarchy, we will use the specificity to make a new term hierarchy or to augment terms to the existing term hierarchy. We will also examine the balance of the existing term hierarchy using term specificity.

References

[1] A. Aizawa, An information-theoretic perspective of tf-idf measures, Journal of Information Processing and Management 39 (2003).
[2] S.A. Caraballo, E. Charniak, Determining the specificity of nouns from text, in: Proceedings of the Joint SIGDAT Conference on EMNLP and Very Large Corpora, 1999.
[3] T.M. Cover, J.A. Tomas, Elements of Information Theory, John Wiley and Sons Inc., New York, 1991.
[4] W. Croft, Typology and Universals, second ed., Cambridge Textbooks in Linguistics, Cambridge University Press, Cambridge, 2004.
[5] K. Frantzi, S. Ananiadou, H. Mima, Automatic recognition of multi-word terms: the C-value/NC-value method, Journal of Digital Libraries 3 (2) (2000).
[6] ISO 704, Terminology work—Principles and methods, ISO 704:2000(E), 2000.
[7] J.C. Sager, Handbook of Term Management, vol. 1, John Benjamin Publishing Company, New York, 1997.
[8] S.K.M. Wong, Y.Y. Yao, An information-theoretic measure of term specificity, Journal of the American Society for Information Science 43 (1) (1992).
[9] R. Forsyth, R. Rada, Adding an edge, in: Machine Learning: Applications in Expert Systems and Information Retrieval, Ellis Horwood Ltd., Chichester, UK, 1986, pp. 198–212.
[10] W.A. Woods, Conceptual indexing: a better way to organize knowledge, Sun Labs Technical Report, TR-97-61.