Domain term relevance through *tf-dcf*

Lucelene Lopes
PPGCC - FACIN
PUCRS University
Porto Alegre - Brazil
lucelene.lopes@pucrs.br

Paulo Fernandes
PPGCC - FACIN
PUCRS University
Porto Alegre - Brazil
paulo.fernandes@pucrs.br

Renata Vieira
PPGCC - FACIN
PUCRS University
Porto Alegre - Brazil
renata.vieira@pucrs.br

***Abstract*—This paper proposes a new index for the relevance of terms extracted from domain corpora. We call it term frequency, disjoint corpora frequency (*tf-dcf*), and it is based on the absolute term frequency of each term tempered by its frequency in other (contrasting) corpora. Conceptual differences and mathematical computation of the proposed index are discussed in respect with other similar approaches that also take the frequency in contrasting corpora into account. To illustrate the efficiency of the *tf-dcf* index, this paper evaluates the application of this index and other similar approaches.**

I. Introduction

The automatic extraction of terms from texts is a well mapped task, but the automatic choice of which extracted terms are relevant for a specific domain is a much more defiant task. Finding the most relevant terms for a domain, *i.e.*, the domain concepts, is an important step for knowledge engineering tasks such as ontology learning from texts [1].

Some classical linguistic-based work in this area suggest the use of distributional analysis [2] to associate terms and then, establish which of them are good concept candidates. A different approach, but yet following the same idea of inferring concepts from term association, is made by Chemudugunta *et al.* [3], where the identification of concepts is made through pure statistical measures tempered by previous inserted human information. Titov and Kozhevnikov [4] work also follows this line of research by inferring semantic relations among terms in order to identify different terms representing a same concept in sets of small documents (weather forecasts) with no linguistic annotation.

The work of Bosma and Vossen [5] presents a similar effort to establish term relevance measures considering a multiple corpora resource. This work proposes different relevance measures of terms to each corpus, but, Bosma and Vossen's relevance measure of a term in a given corpus do not affect the relevance of this same term in other corpora. In fact, the methodology proposed in their work access WORDNET [6] in order to validate the term candidates according to their measures, but also to establish relations (hypernym, hyponym, meronym, *etc.*) among them.

In opposition to these efforts, this paper proposes an approach that is not linguistic-based, but it relies only the on statistical information gather from the domain corpus to establish a numerical measure to term relevance in this corpus. Therefore, this paper approach is aligned with works that take into account the term frequency on documents to compute a relevance index to establish how representative a term extracted from a corpus will be for the domain represented by this corpus. Some examples of such statistical-based approaches are the works of Dunning in 1993 [7] which proposes the use of log likelihood ratio, Manning and Schultz in 1999 [8] which proposes a composition of *tf-idf* (term-frequency, inverse document frequency [9] adapted for term relevance in a corpus), and other initiatives based on computing indexes from one specific corpus only.

However, our claim is that those typical indices fail to rule out those terms which are not particularly relevant to a target domain. The basic idea behind approaches like the one in our paper is the assumption that a term relevance to a specific domain can only be established by comparison with corpora from other domains, called contrasting corpora.

One of the first examples of similar previous work like our own was the work of Chung in 2003 [10]. But recently, more sophisticated versions were proposed by Park *et al.* in 2008 [11] with domain specificity index, by Kit and Liu in 2008 with termhood index [12], and by Kim *et al.* in 2009 with term frequency, inverse domain frequency index [13]. These approaches brought some quality to the term extraction, as was verified by the works of Teixeira *et al.* [14], as well as, Rose *et al.* [15].

Similar to our proposal, all these previous works followed the same principle to compute a relevance index that is directly proportional to the term absolute frequency in the corpus and inversely proportional to the term absolute frequency in other corpora. The main difference among these similar previous works [11], [12], [13] and our own is the specific formula to weight the influence of other corpora frequency.

This paper first contribution resides in drawing a panorama of options of indices to express the relevance of extracted term from a domain corpus, focusing on indices that take into account also corpora of other domains (contrasting corpora). Some experiments illustrate the benefits of approaches using contrasting corpora over traditional indices.

Secondly, and most important, this paper contributes with the proposal of a new relevance index, called *tf-dcf*, that is, according to our experiments, superior to the other indexes based on contrasting corpora. This contribution is enhanced by the analysis of the *tf-dcf* behavior against different options of contrasting corpora.

It is not the goal of this paper to analyze techniques to improve the quality of term extraction itself, since we assume that a previously performed extraction provides a set of extracted terms. It is also out of the scope of this paper to analyze how many terms should be considered concepts of a domain. Our purpose is to present arguments and experiments showing that the proposed index is effective to rank extracted terms according to their relevance for the domain, thus allowing to identify domain concept candidates.

This paper is organized as follows: Section II describes the existent statistical measures that are compared to our proposed *tf-dcf* index; Section III presents our paper main contribution, which is the proposal of a term relevance index based on the inclusion of a "disjoint corpora frequency" (*dcf*) component; Section IV evaluates the existing and proposed indices. Finally, Conclusion stress the contributions and limitations of this paper, leading to the proposition of future works.

II. Existing measures for relevance estimation

The most elementary way to establish the statistical relevance of terms extracted from a domain specific corpus is to compute the absolute frequency of terms, *i.e.*, how many times each term occurs in the corpus. Obviously, this simple approach is very fragile, since not necessarily a very frequent term is relevant for the domain. This fact is specially noticeable with simple extraction methods, although even sophisticated linguistic-based methods also suffer from using such simple criteria.

For example, pure statistical methods require the adoption of a list of highly frequent grammatical words (*stop list*). Without a stop list, any pure statistical method delivers terms with very low significance such as prepositions and usual expressions. However, it might be very difficult to establish an exhaustive stop list in advance for different domain and genre.

The use of term frequency as relevance measure is a little less harmful for extraction methods taking into account linguistic information. For example, the syntactic annotation of a corpus allows the extraction procedure to avoid terms that are unsuitable for concept names, such as verbs and pronouns. In fact, more sophisticated linguistic analysis, as the identification of noun phrases, may improve significantly the quality of extraction, but even in these cases the use of term frequency do not prevent the incorrect extraction of common expressions which are not domain specific. For example, the quite common expression "*future work*" may be found in several academic texts, but it is hardly considered a defining concept to any scientific domain.

Nevertheless, the starting point of all sophisticated indices is the simple absolute frequency. Assuming, $tf_{t,d}$ as the number of occurrences of term t in document d, and $\mathcal{D}^{(c)}$ the set of all documents belonging to the corpus c referring to a specific domain, the absolute term frequency of a term t in corpus c is expressed by:

$$tf_t^{(c)} = \sum_{\forall d \in \mathcal{D}^{(c)}} tf_{t,d} \tag{1}$$

A. Term frequency and inverse document frequency - tf-idf

An alternative for plain term frequency is to take into account the frequency of the term among documents. The seminal work of Spärck-Jones [9] shows the importance to consider frequent terms, but also non-frequent ones in order to retrieve documents. These ideas lead to the well-known Robertson and Spärck-Jones probabilistic model to term relevance to specific documents [16]. Croft and Harper [17], and later Robertson and Walker [18], proposed formulations to a popular index that takes positively into account the term frequency (*tf*), *i.e.*, the number of occurrences of a given term t in a document d; and also considers negatively the number of documents of the corpus where term t appears at least once, *i.e.*, the inverse document frequency (*idf*).

This index, called *tf-idf* has many formulations, *e.g.*, [19], [20], [8], but in this paper we will consider the formulation adopted by Bell *et al.* [21]. The *tf-idf* index is mathematically defined for each term t to each document d belonging to a corpus c that has at least one occurrence of t as follows:

$$tf\text{-}idf_{t,d} = \underbrace{(1 + \log(tf_{t,d}))}_{tf \text{ part}} \times \underbrace{\log\left(1 + \frac{\mid \mathcal{D}^{(c)} \mid}{\mid \mathcal{D}_t^{(c)} \mid}\right)}_{idf \text{ part}} \tag{2}$$

where $tf_{t,d}$ is the number of occurrences of term t in document d; $\mathcal{D}^{(c)}$ is the set of all document of a given corpus c; and $\mathcal{D}_t^{(c)}$ is the subset of these documents where t appears at least once.

Observing equation (2) it is possible to observe the term frequency (*tf*) and the inverse document frequency (*idf*) parts. The *tf* part considers the logarithmic frequency of the term, since the variation of term occurrences of terms approaches an exponential distribution, *i.e.*, a term t that occurs 10 times is not 10 times more important than a term t' that appears only once. Nevertheless, term t is an order of magnitude more important than term t'. The *idf* part represents a value that varies from $\log(2)$ for a term that appears in all documents, until $\log(1+ \mid \mathcal{D}^{(c)} \mid)$ for a document that appears in only one document.

The idea behind *tf-idf* formulation is that a term t is more relevant as a keyword for a document d if it appears many times in this document and very few times (or ideally none) in other documents. This is an important distinction for information retrieval. The popularity of this index is justified mostly because it prevents frequent terms spread in many documents to be considered more relevant than they should. Indeed, *tf-idf* is an effective measure to identify the defining terms of documents, because it spots terms that are good for document indexation.

The use of *tf-idf* to establish relevance of terms to domain corpora was proposed by Manning and Schütze [8]. According to these authors, a possible index to express the relevance of a term t in a corpus c is expressed by:

$$tf\text{-}idf_t^{(c)} = \sum_{\forall d \in \mathcal{D}_t^{(c)}} tf\text{-}idf_{t,d} \tag{3}$$

B. Term domain specificity - tds

The first initiatives to consider the relevance of terms to a domain corpus taking into account contrastive generic corpus, or corpora, include the works made by Chung in 2003 [10] and Drouin in 2004 [22]. However, at the authors best knowledge, it is the work of Park *et al.* [11], in 2008, one of the first formulations of an index to express term relevance to a specific domain. In that work, such index is called *domain specificity*, and it is expressed as the ratio between the probability of occurrence of a term t in a domain corpus c and the probability of this same term in a generic corpus. Park *et al.* definition of term t domain specificity to a specific domain corpus c, considering a generic domain corpus g was expressed as:

$$tds_t^{(c)} = \frac{p_t^{(c)}}{p_t^{(g)}} = \frac{\frac{tf_t^{(c)}}{N^{(c)}}}{\frac{tf_t^{(g)}}{N^{(g)}}} \tag{4}$$

where $p_t^{(c)}$ express the probability of occurrence of term t in corpus c; and $N^{(c)}$ is the total number of terms in corpus c, *i.e.*, $N^{(c)} = \sum_{\forall t'} tf_{t'}^{(c)}$.

C. Termhood - thd

Following the approach to consider, besides the domain corpus of interest, a contrasting corpus, the work of Kit and Liu in 2008 [12] proposes an index called termhood. This index, as for Park *et al.*'s term domain specificity, follows the idea that a term relevant to a domain is more frequent in the corpus domain than in other corpora. The main difference brought by this work is to consider the term rank in the corpus vocabulary (the set of all terms in the corpus), instead of the term absolute frequency. Kit and Liu definition of term t termhood index for a corpus c, a generic domain corpus g (called background corpus by them) was expressed by:

$$thd_t^{(c)} = \underbrace{\frac{r_t^{(c)}}{\mid V^{(c)} \mid}}_{\text{norm. rank value in } c} - \underbrace{\frac{r_t^{(g)}}{\mid V^{(g)} \mid}}_{\text{norm. rank value in } g} \tag{5}$$

where $V^{(c)}$ is the vocabulary of corpus c, *i.e.*, $\mid V^{(c)} \mid$ is the cardinality of the set of all terms in the corpus c, and $r_t^{(c)}$ is the rank value of term t expressed as $\mid V^{(c)} \mid$ for the more frequent term, $\mid V^{(c)} \mid -1$ for the second most frequent, and so on until the less frequent term as $r_t^{(c)} = 1$.

Observing the termhood index we can see it as the difference between the normalized rank value of the term in the domain corpus c and the generic domain corpus g. Actually, the division of the rank value by the vocabulary size is intended to keep the normalized rank value within the interval $(0, 1]$, with a value equal to 1 to the more frequent term, and the other terms decaying, according to their frequency, asymptotically toward 0.

As a result, the termhood index will be whitin the interval $[1, -1]$, having the more frequent term in c having a value equal to 1, if it does not belong to vocabulary $V^{(g)}$, until a value -1 for the more frequent term in g, if it does not belong to vocabulary $V^{(c)}$.

D. Term frequency, inverse domain frequency - TF-IDF

Recently, Kim *et al.* [13] have proposed in 2009 another index to rank term relevance considering the original idea of the *tf-idf* index, which was to identify whereas a term is suitable to represent a document. In such way, Kim *et al.* did not actually proposed a new index, but instead, they proposed the use of the same *tf-idf* formulation, but considering the set of documents of a corpus as a single document. To avoid confusion, we will refer to this index with the acronym *TF-IDF* in uppercase, to differentiate it from the term frequency, inverse document frequency (*tf-idf*).

The *TF-IDF* index for term t at corpus c, considering a set of corpora $\mathcal{G}$ as proposed by Kim *et al.* is numerically expressed by:

$$\textit{TF-IDF}_t^{(c)} = \underbrace{\frac{tf_t^{(c)}}{\sum_{\forall t'} tf_{t'}^{(c)}}}_{\textit{TF} \text{ part}} \times \underbrace{\log\left(\frac{\mid \mathcal{G} \mid}{\mid \mathcal{G}_t \mid}\right)}_{\textit{IDF} \text{ part}} \tag{6}$$

where $tf_t^{(c)}$ is the term frequency of term t in corpus c; $\mathcal{G}$ is the set of all domain corpora; and $\mathcal{G}_t$ is the subset of $\mathcal{G}$ where the term t appears at least once.

It is important to notice that the basic formulation of *tf-idf* used as inspiration by Kim *et al.* proposal is not as robust as the one of Bell *et al.* (Eq. 3). For instance, if a term t appears in all corpora, the *IDF* part of Eq. 6 will become 0, and therefore, such term t will have a *TF-IDF* index also equal to 0, *i.e.*, it will be considered less relevant than any other term, regardless its number of occurrences. Another important difference between Equations 3 and 6 is that Bell *et al.*'s (Eq. 3) uses the log of absolute term frequency in the *tf* part, while Kim *et al.*'s (Eq. 6) considers directly a relative term frequency.

III. Proposed Index

The goal of all indices presented in the previous section is to obtain higher numeric values for terms that are relevant to a given domain, or for more recent knowledge engineering tasks [14], [15], terms that are suitable candidates for concepts of an ontology. The raw term absolute frequency (Eq. 1), obviously indicates a relevance, since a term that is very frequent is likely to be important to the domain. Also the *tf-idf* (Eq. 3) index can be an indicative of relevance, since terms that are very distinctive to some documents of the corpus are also likely to be representative of the domain.

The *tds* (Eq. 4), *thd* (Eq. 5) and *TF-IDF* (Eq. 6) indices have better chance to identifying concepts of a domain because they use contrasting corpora. Nevertheless, these indices adopt different approaches that reveals distinct empirical initiatives to tackle the concept identification problem.

The first difference is how these indices take the occurrences of terms in the domain corpus into account. The *tds* (Eq. 4) and *TF-IDF* (Eq. 6) indices compute a relative frequency of the term, since the term probability ($p_t^{(c)}$) for *tds* and the *tf* part for *TF-IDF* are computed as the absolute frequency divided

by the total number of terms in the domain corpus. The *thd* (Eq. 5) index, however, computes a normalized rank value, that, even though being computed according to the absolute frequency, delivers a linear relation[1] among all terms.

The second difference resides in the effect brought by the occurrence of terms in contrasting corpora. The *tds* (Eq. 4) index penalizes the terms that occurs in the contrasting corpora by dividing its probability in the domain corpus by the probability in the contrasting corpora. The *thd* (Eq. 5) index also penalizes the terms that occurs in the contrasting corpora, but in this case it subtracts the normalized rank value in the domain corpus by the normalized rank value in the contrasting corpora. The approach for *TF-IDF* (Eq. 6) index is quite different, since it rewards the terms that are unique to the domain corpus by multiplying the relative frequency by the log of the number of corpora. Such reward decreases as the term appears in other contrasting corpora, until it drops to 0 when the term appears in all corpora. It is important to notice that this reward decreases proportionally to the number of corpora, but it is independent to the number of term occurrences in contrasting corpora.

We propose a new index to estimate the term relevance to a domain following the same idea of contrasting corpora, but we propose differences in the way term occurrences in the domain corpus are taken into account, and most of all, in the effect brought by occurrences in the contrasting corpora. Specifically, we propose a representation to this effect called "disjoint corpora frequency" (*dcf*), which is a mathematical way to penalize terms that appear in contrasting corpora proportionally to its number of occurrences, as well as the number of contrasting corpora in which the term appears.

A. Term frequency, disjoint corpora frequency - tf-dcf

Our proposal, like other contrasting corpora approaches, is based on a primary indication of term relevance and a reward/penalization mechanism. The basis of *tf-dcf* index is to consider the absolute frequency as the primary indication of term relevance. Then, we choose to penalize terms that appear in the contrasting corpora by dividing its absolute frequency in the domain corpus by a geometric composition of its absolute frequency in each of the contrasting corpora. The *tf-dcf* index is mathematically expressed, for term t in corpus c, considering a set of contrasting corpora $\mathcal{G}$, as:

$$\textit{tf-dcf}_t^{(c)} = \frac{\textit{tf}_t^{(c)}}{\prod\limits_{\forall g \in \mathcal{G}} 1 + \log\left(1 + \textit{tf}_t^{(g)}\right)} \tag{7}$$

The choice of absolute frequency as primary indication of term t relevance for corpus c, instead of using a relative frequency (like *tds* and *TF-IDF*) or term rank (like *thd*), aims the simplicity of the measure for two main reasons:

- We do not consider that there is a need for linearization brought by the use of the term rank, as for *thd* index, nor there is a need to make explicit the normalization according to the corpus size, as for *tds* and *TF-IDF*; In fact, any normalization according to the corpus size still remain possible after the *tf-dcf* computation;
- We consider that keeping a relation with the absolute term frequency preserves the index intuitive comprehension, since the *tf-dcf* index numeric value will be smaller (if the term appears in the contrasting corpora) or equal to *tf* (if the term does not appear in the contrasting corpora).

The geometric composition of absolute frequencies in the contrasting corpora chosen to express the penalization, *i.e.*, the divisor in Eq. 7, tries to encompass the following assumptions:

- The number of occurrences of a term in each of the contrasting corpora is distributed according to a Zipf law [23], and to correctly estimated this importance, a linearization of this number of occurrences must be made;
- A term that appears only in the domain corpora should not be penalized at all, *i.e.*, terms that do not occur in the contrasting corpora must have the divisor equal to 1; and
- A term that appears in many corpora is more likely to be irrelevant to the domain corpus, than those terms that appears in fewer corpora.

Because of the first assumption, we choose to consider a log function to compute the absolute frequency in each contrasting corpora ($\textit{tf}_t^{(g)}$). This decision follows the same principle adopted in the original proposition of *tf-idf* measure proposed by Robertson and Spärck-Jones [16].

The second assumption made us adapt this log function with the addition of value 1 inside and outside the log function in order to deliver a value equal to 1 when the number of occurrences of a term in a contrasting corpora is equal to 0. This decision follows the same principle adopted to the Bell *et al.* [21] to express their formulation of *tf-idf* measure.

Finally, the third assumption led us to employ the product of the log of occurrences in each contrasting corpora. The product represents that the importance of occurrences grows geometrically as it appears in other corpora. In fact, according to our formulation a term is more likely to be irrelevant for a domain corpus when it appears few times in many multiple contrasting corpora, than if it appears many times in just few contrasting corpora. Additionally, the product is compatible with the idea to have a divisor equal to 1 when a term appears only in the domain corpus.

IV. Practical Results

The practical application of the proposed index is meant to illustrate its effectiveness and some basic characteristics of *tf-dcf* according to the contrasting corpora used. The experiments were conducted over Brazilian Portuguese corpora, using a linguistic-based term extraction tool to provide terms and their number of occurrences. Nevertheless, corpora in any language submitted to any kind of extraction could be employed without any loss of generality.

[1]It is important to recall, that the distribution of absolute frequency values is likely to follow a Zipf law [23], *i.e.*, the most frequent term is likely to have twice the number of occurrences as the second, three times the number of occurrences of the third, and so on.

A. The chosen corpora

The chosen test bed was one corpus from Pediatrics domain [24] with 281 documents from The Brazilian Journal on Pediatrics. This corpus (PED) was chosen because of the availability of reference lists of relevant terms.

Four other scientific corpora were used as support for definition of specific Pediatrics terms. These corpora have approximatively 1 million words each and their domains are: Stochastic modeling (SM), Data mining (DM), Parallel processing (PP) and Geology (GEO) [25]. Tab. I summarizes the information about these corpora.

Table I
CORPORA CHARACTERISTICS.

		documents	sentences	words
Pediatrics	PED	281	27,724	835,412
Stochastic Modeling	SM	88	44,222	1,173,401
Data Mining	DM	53	42,932	1,127,816
Parallel Processing	PP	62	40,928	1,086,771
Geology	GEO	234	69,461	2,010,527

B. Extraction tools

The extraction procedure of terms and their frequencies was made by a two step process. First the documents were annotated by the Portuguese parser PALAVRAS [26]. Then the PALAVRAS output, *i.e.*, a set of TigerXML files, was submitted to ExATOlp term extractor [27].

PALAVRAS and ExATOlp joint application delivers high quality term lists, since the extracted terms are noun phrases found in the corpus and their frequencies. The extracted noun phrases were filtered according to ExATOlp heuristic rules aiming the output of noun phrases as meaningful as possible. These heuristics goes from simple exclusion of articles, but also quite ingenious ones like detection of implicit noun phrases[2] [28].

C. Extracted terms and reference lists

The extracted terms were divided in two lists, bigrams and trigrams. Single terms and those with more than three words were not considered in the evaluation, since they were not included in the hand-made reference list constructed by terminology laboratory TEXTECC (http://www6.ufrgs.br/textecc/).

The reference lists were produced by a careful and laborious process that involved terminologists, domain specialists (Pediatricians) and academic students. These lists are available for download at TEXTECC website and they have been used for practical applications including glossary construction, translation aid, and even ontology construction. These reference lists are composed by 1,534 bigrams and 2,660 trigrams and they can also be consulted at http://ontolp.inf.pucrs.br/ontolp/downloads-ontolplista.php.

The full extracted term lists delivered by PALAVRAS and ExATOlp for the Pediatrics corpus were composed by 15,483 distinct bigrams and 18,171 distinct trigrams. To each of these lists the computed indices were:

- *tf* the absolute term frequency (Eq. 1);
- *tf-idf* the term frequency, inverse document frequency (Eq. 3) with the basic formulation from Bell *et al.* [21] aggregated with the sum proposed by Manning and Schütze [8] to be used as an example of index not using contrasting corpora;
- *tds* the term domain specificity (Eq. 4) proposed by Park *et al.* [11];
- *thd* the termhood (Eq. 5) proposed by Kit and Liu [12];
- *TF-IDF* the term frequency, inverse domain frequency (Eq. 6) proposed by Kim *et al.* [13]; and
- *tf-dcf* the term frequency, disjoint corpora frequency (Eq. 7) proposed in the previous section of this paper.

D. The impact of different measures on frequent terms

Observing in detail some terms in the extracted lists it is possible to have a better understanding of the effect of each index, and, therefore, the benefits brought by *tf-dcf* as relevance index. Tab. II presents the top ten frequent terms, *i.e.*, the ten terms with more absolute occurrences in the Pediatrics corpus. In this table it is shown the number of occurrences of the term in each corpora, *i.e.*, Pediatrics (PED), Stochastic modeling (SM), Data mining (DM), Parallel processing (PP) and Geology (GEO). Additionally, the last column (ref. list) indicates wether the term belongs ("IN") or not ("OUT") to the reference list.

Table II
OCCURRENCES FOR FREQUENT TERMS FROM PEDIATRICS CORPUS.

term in Portuguese	(translation)	PED	SM	DM	PP	GEO	ref. list
aleitamento materno	(breast feeding)	306	0	0	0	0	IN
recém nascido	(new born)	299	0	0	0	0	IN
faixa etária	(age slot)	234	0	6	0	0	IN
presente estudo	(current study)	188	4	1	0	67	OUT
leite materno	(mother's milk)	163	0	0	0	0	IN
idade gestacional	(gestacional age)	144	0	0	0	0	IN
ventilação mecânica	(mechanical ventilation)	138	0	0	0	0	IN
via aérea	(airway)	120	0	0	0	0	IN
pressão arterial	(blood pressure)	112	0	0	0	0	IN
sexo masculino	(male sex)	109	7	8	0	0	OUT

The same ten more frequent terms are also shown in Tab. III with the values for the six presented indices, as well as their rank according to each of them. For example, in the third row of Tab. III, the term "faixa etária" ("age slot" in English) belongs to the reference list and it is ranked as the third term in the lists sorted with the term frequency (*tf* - Eq. 1) and with the term frequency, inverse document frequency (*tf-idf* - Eq. 3). In the lists sorted with the other indices this term is ranked as the $13{,}281^{th}$ (for *tds* - Eq. 4), the fourth (for *thd* - Eq. 5), the sixth (for *TF-IDF* - Eq. 6), and the fifteenth (for *tf-dcf* - Eq. 7).

Observing the rank differences between the lists sorted with the term frequency (*tf* - Eq. 1) and the term frequency, inverse document frequency (*tf-idf* - Eq. 3), we noticed an important

[2]Implicit noun phrases are, for example, "sick children" and "healthy children" that can be extracted from the sentence "Sick and healthy children can be treated.".

Table III
ANALYSIS OF FREQUENT TERMS FROM PEDIATRICS CORPUS.

term in Portuguese (translation)	*tf* Eq. 1	*tf-idf* Eq. 3	*tds* Eq. 4	*thd* Eq. 5	*TF-IDF* Eq. 6	*tf-dcf* Eq. 7
aleitamento materno (breast feeding)	306 1^{st}	199.18 1^{st}	1.00 1^{st}	1.00 1^{st}	0.0027 1^{st}	306.00 1^{th}
recém nascido (new born)	299 2^{nd}	184.98 2^{nd}	1.00 1^{st}	0.99 2^{nd}	0.0027 2^{nd}	299.00 2^{nd}
faixa etária (age slot)	234 3^{rd}	169.18 3^{rd}	0.98 $13{,}281^{st}$	0.93 4^{th}	0.0012 6^{th}	61.46 15^{th}
presente estudo (current study)	188 4^{th}	167.78 4^{th}	0.73 $13{,}429^{th}$	0.50 42^{nd}	0.0002 57^{th}	3.99 $1{,}276^{th}$
leite materno (mother's milk)	163 5^{th}	143.23 5^{th}	1.00 1^{st}	0.94 3^{rd}	0.0015 3^{rd}	163.00 3^{rd}
idade gestacional (gestacional age)	144 6^{th}	135.60 7^{th}	1.00 1^{st}	0.93 5^{th}	0.0013 4^{th}	144.00 4^{th}
ventilação mecânica (mechanical ventilation)	138 7^{th}	140.85 6^{th}	1.00 1^{st}	0.91 6^{th}	0.0012 5^{th}	138.00 5^{th}
via aérea (airway)	120 8^{th}	132.72 8^{th}	1.00 1^{st}	0.90 7^{th}	0.0011 7^{th}	120.00 6^{th}
pressão arterial (blood pressure)	112 9^{th}	93.27 19^{th}	1.00 1^{st}	0.88 8^{th}	0.0010 8^{th}	112.00 7^{th}
sexo masculino (male sex)	109 10^{th}	125.70 9^{th}	0.88 $13{,}318^{th}$	0.77 14^{th}	0.0003 35^{th}	6.53 543^{th}

similarity. The only significantly change occurs for the term "pressão arterial" ("blood pressure") that drops from the 9^{th} to the 19^{th} position. However, this change does not correspond to a meaningful downgrade, since this term ("blood pressure") seems to be as relevant to Pediatrics as, for instance, "via aérea" ("airway"). In contrast, the quite generic term "presente estudo" ("current study") is not affected at all by *tf-idf*.

Observing the effect brought by the term domain specificity index (*tds* - Eq. 4), we realize the lack of precision, since it assigns an equally important rank to all terms that are not exclusive to the Pediatrics corpus. Consequently, the terms that appears in other corpora are cast out of any list of relevant terms, since, giving the contrasting corpora (SM, DM, PP and GEO), there is more than 13,000 terms appearing only in the Pediatrics corpus. The terms "faixa etária" ("age slot"), "presente estudo" ("current study") and "sexo masculino" ("male sex") are all ranked beyond the $13{,}000^{th}$ position.

The list sorted with the termhood index (*thd* - Eq. 5) shows the downgrade effect on the three terms appearing in the contrasting corpora (grey rows in Tabs. II and III). However, these terms are not sent very low, since even the term "presente estudo" ("current study"), which is very frequent in the contrasting corpora (72 occurrences), is downgraded only to the 42^{th} position.

The list sorted according to term frequency, inverse domain frequency index (*TF-IDF* - Eq. 6) shows a stronger effect than the termhood (*thd* - Eq. 5), since it is based on the number of contrasting corpora the term appear. In consequence, the term "faixa etária" ("age slot") drops to the sixth position because it appears also in the Data Mining corpus, while the term "presente estudo" ("current study") drops to the 57^{th} position because it appears in all corpora, but Geology.

It is important to call the reader attention that our proposed index (*tf-dcf* - Eq. 7) is the only one that takes into account both the number of occurrences in the contrasting corpora (as termhood and term domain specificity), and the number of corpora in which the term appears (as term frequency, inverse corpus frequency). For that reason, the downgrade effect in the list sorted according to our index is the stronger one. Our index casts out the term "presente estudo" ("current study") to the $1{,}276^{th}$ position, while it downgrades significantly the term "sexo masculino" ("male sex") to the 543^{th} position. In opposition, the term "faixa etária" ("age slot") is mildly downgraded from the third to the fifteenth position.

V. CONCLUSION

This paper presented a novel numerical index to estimate the relevance of extracted terms with respect to a specific domain. The inclusion of disjoint corpora frequency (*dcf*) component successfully improved the precision of extracted lists in comparison with the traditional *tf* and *tf-idf*, but also other indices based on comparison with contrasting corpora, namely term domain specificity [11], termhood [12] and term frequency, inverse domain frequency [13].

The proposed *dcf* approach was described here in composition with the absolute frequency (*tf*) and it has the advantage to keep an analogue semantic of the original absolute frequency index. If a given term does not appear in other corpora, its *tf-dcf* index will be equal to the term frequency, *i.e.*, only terms appearing in other corpora will be numerically downgraded. This is not the case of any of the other pre-existent measures.

Our proposal is the follow up to initial studies based on the comparison with contrasting corpora. Such intuitive idea was initially proposed during the last 10 years [10], [22], [29], [11], [12], [13], [15], but, at the authors best knowledge, our proposal is the first one to pay attention to an correct weighting of the influence of occurrences of terms in contrasting corpora.

Specifically, our *tf-dcf* index formulation consider the product of the log of the number of occurrences in other corpora as reductive factor for the domain corpus absolute term frequency. This choice is justified by the fact that term occurrences are likely to be distributed by a Zipf law [23]. In Park *et al.* [11] this fact was ignored. In Kit and Liu [12] this fact was approached by the rank difference. In Kim *et al.* [13] this fact was approached by term relative frequency and the logarithm in the *IDF* part. Therefore, our formulation seems to be mathematically more robust.

The main limitation of the current study is the lack of thorough experiments with other corpora. We had choose to limit our experiments to the studied corpora because there were no sign of availability of data sets previously employed by other authors. Nevertheless, since the objective of this paper is to propose the *tf-dcf* index, it remains as a natural future work the experimentation of our proposal to a statistically significant set of corpora. Such future work will demand the analysis of the proposed *tf-dcf* index, in comparison with other indices, in terms of numerical measures, as precision, and the gathering of corpora and corresponding lists of references.

Another valid future work is the study of heuristics to choose a good cut-off point to apply in the extracted term lists. With the use of a simple index of relevance, like the absolute term frequency, the cut-off point choice seems simple, since it is enough to define a minimum number of term occurrences. However, with a more sophisticated one, as the *tf-dcf* index proposed here, it is a little less obvious to define a meaningful and effective cut-off point [30].

REFERENCES

[1] P. Cimiano, *Ontology learning and population from text: algorithms, evaluation and applications*. Springer, 2006.

[2] D. Bourigault and G. Lame, "Analyse distributionnelle et structuration de terminologie. application a la construction d'une ontologie documentaire du droit," *Traitement automatique des langues*, vol. 43, no. 1, 2002.

[3] C. Chemudugunta, A. Holloway, P. Smyth, and M. Steyvers, "Modeling documents by combining semantic concepts with unsupervised statistical learning," in *The Semantic Web - ISWC 2008*, ser. Lecture Notes in Computer Science, A. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, and K. Thirunarayan, Eds. Springer Berlin / Heidelberg, 2008, vol. 5318, pp. 229–244.

[4] I. Titov and M. Kozhevnikov, "Bootstrapping semantic analyzers from non-contradictory texts," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ser. ACL '10. Morristown, NJ, USA: Association for Computational Linguistics, 2010, pp. 958–967.

[5] W. Bosma and P. Vossen, "Bootstrapping language neutral term extraction," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, Eds. Valletta, Malta: European Language Resources Association (ELRA), may 2010.

[6] C. Fellbaum, "Wordnet," in *Theory and Applications of Ontology: Computer Applications*, R. Poli, M. Healy, and A. Kameas, Eds. Springer Netherlands, 2010, pp. 231–243.

[7] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*, vol. 19, pp. 61–74, March 1993. [Online]. Available: http://dl.acm.org/citation.cfm?id=972450.972454

[8] C. D. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT Press, 1999.

[9] K. Spärck-Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972. [Online]. Available: http://www.emeraldinsight.com/journals.htm?articleid=1649768&show=abstract

[10] T. Chung, "A corpus comparison approach for terminology extraction," *Terminology*, vol. 9, pp. 221–246, 2003. [Online]. Available: http://www.ingentaconnect.com/content/jbp/term/2003/00000009/00000002/art00004

[11] Y. Park, S. Patwardhan, K. Visweswariah, and S. C. Gates, "An empirical analysis of word error rate and keyword error rate," in *INTERSPEECH*, 2008, pp. 2070–2073.

[12] C. Kit and X. Liu, "Measuring mono-word termhood by rank difference via corpus comparison," *Terminology*, vol. 14, no. 2, pp. 204–229, 2008.

[13] S. N. Kim, T. Baldwin, and M.-Y. Kan, "Extracting domain-specific words - a statistical approach," in *Proceedings of the 2009 Australasian Language Technology Association Workshop*, L. Pizzato and R. Schwitter, Eds. Sydney, Australia: Australasian Language Technology Association, December 2009, pp. 94–98. [Online]. Available: www.alta.asn.au/events/alta2009/proceedings/pdf/ALTA2009_12.pdf

[14] L. Teixeira, G. Lopes, and R. Ribeiro, "Automatic extraction of document topics," in *Technological Innovation for Sustainability*, ser. IFIP Advances in Information and Communication Technology, L. Camarinha-Matos, Ed. Springer Boston, 2011, vol. 349, pp. 101–108. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-19170-1_11

[15] G. Rose, M. Holland, S. Larocca, and R. Winkler, "Semi-automated methods for refining a domain-specific terminology base," U. S. Army Research Laboratory, Adelphi, MD, USA, Tech. Rep. ARL-RP-0311, 2011.

[16] S. Robertson and K. Spärck-Jones, "Relevance weighting of search terms," *Journal of American Society for Information Science*, vol. 27, no. 3, pp. 129–146, 1976.

[17] W. B. Croft and D. J. Harper, "Using probabilistic models of document retrieval without relevance information," *Journal of documentation*, vol. 35, no. 4, pp. 285–295, 1979.

[18] S. E. Robertson and S. Walker, "On relevance weights with little relevance information," *SIGIR Forum*, vol. 31, pp. 16–24, July 1997. [Online]. Available: http://doi.acm.org/10.1145/278459.258529

[19] A. Lavelli, F. Sebastiani, and R. Zanoli, "Distributional term representations: an experimental comparison," in *CIKM*, 2004, pp. 615–624.

[20] A. Maedche and S. Staab, "Learning ontologies for the semantic web," in *SemWeb*, 2001.

[21] T. Bell, I. Witten, and A. Moffat, *Managing Gigabytes: Compressing and Indexing Documents and Images*. San Francisco: Morgan Kaufmann, 1999. [Online]. Available: http://ontology.csse.uwa.edu.au/reference/browse_paper.php?pid=233281449

[22] P. Drouin, "Detection of domain specific terminology using corpora comparison," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC) 2004*, M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva, Eds., ELRA. Lisbon, Portugal: European Language Resources Association, May 2004, pp. 79–82.

[23] G. K. Zipf, *The Psycho-Biology of Language - An Introduction to Dynamic Philology*. Boston, USA: Houghton-Mifflin Company, 1935.

[24] R. J. Coulthard, "The application of Corpus Methodology to Translation: the JPED parallel corpus and the Pediatrics comparable corpus," Ph.D. dissertation, UFSC, 2005.

[25] L. Lopes and R. Vieira, "Building Domain Specific Corpora in Portuguese Language," Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, Brasil, Tech. Rep. TR 062, Dezembro 2010.

[26] E. Bick, "The parsing system PALAVRAS: automatic grammatical analysis of portuguese in constraint grammar framework," Ph.D. dissertation, Arhus University, 2000.

[27] L. Lopes, P. Fernandes, R. Vieira, and G. Fedrizzi, "ExATO lp – An Automatic Tool for Term Extraction from Portuguese Language Corpora," in *Proceedings of the 4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC '09)*. Faculty of Mathematics and Computer Science of Adam Mickiewicz University, November 2009, pp. 427–431.

[28] L. Lopes and R. Vieira, "Heuristics to improve ontology term extraction," in *PROPOR 2012 – International Conference on Computational Processing of Portuguese Language*, 2012, submitted.

[29] J. Wermter and U. Hahn, "You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ser. ACL-44. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 785–792.

[30] L. Lopes, R. Vieira, M. Finatto, and D. Martins, "Extracting compound terms from domain corpora," *Journal of the Brazilian Computer Society*, vol. 16, pp. 247–259, 2010, 10.1007/s13173-010-0020-4. [Online]. Available: http://dx.doi.org/10.1007/s13173-010-0020-4