Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

Building and Applying Profiles Through Term Extraction

Lucelene Lopes, Renata Vieira

Computer Science Department – PUCRS University
Porto Alegre – Brazil

{lucelene.lopes,renata.vieira}@pucrs.br

Abstract. *This paper proposes a technique to build entity profiles starting from a set of defining corpora, i.e., a corpus considered as the definition of each entity. The proposed technique is applied in a classification task in order to determine how much a text, or corpus, is related to each of the profiled entities. This technique is general enough to be applied to any kind of entity, however, this paper experiments are conduct over entities describing a set of professors of a computer science graduate school through their advised M.Sc. thesis and Ph.D. dissertations. The profiles of each entity are applied to categorize other texts into one of the builded profiles. The analysis of the obtained results illustrates the power of the proposed technique.*

1. Introduction

The amount of available written material is larger than ever, and it clearly tends to keep growing as not only new material is made available, but also previously produced material is being digitalized and made accessible through the Internet. Often the search for information tends to find as obstacle not the unavailability of texts, but the impossibility to read all available material. In such abundant data environment, the challenge is to automatically gather information from text sources [Balog et al. 2013].

The focus of this paper is to gather information in order to profile entities considering the existence of written material characterizing these entities [Zhou and Chang 2013]. Once these entities are dully profiled, many applications of the profiles may be envisaged [Liu and Fang 2012].

Therefore, this paper objective is to proposed a technique to profile entities according to defining corpora, *i.e.*, a corpus capable to characterize each entity. Additionally, we exemplify the application of such entities profiles to categorize texts according to their great or small similarity to each entity.

Specifically, we chose as entities a group of professors acting on a graduate Computer Science program and we consider as the defining texts of each professor the M.Sc. and Ph.D. dissertations produced under his/her advisory. Therefore, each professor is profiled according to the produced texts under his/her supervision, and these profiles are applied to compute the similarity of other texts to each professor's production, thus allowing to categorize other texts with respect to each professor.

It is important to call the reader attention that the proposed profiling procedure can be applied to any set of entities giving that defining corpora characterizing each entity are available. Also, the exemplified application to categorize texts by the similarity to each entity could be replaced by other applications without any loss of generality.

This paper is organized as follows: the next section briefly presents related work; Section 3 describes the proposed technique to build profiles; Section 4 exemplifies the application of builded profiles to categorize texts; Section 5 presents practical experiments of the proposed technique to a practical case. Finally, the conclusion summarizes this paper contribution and suggests future works.

2. Related Work

Automatic profiling entities is, at the same time, an interesting research topic [Wei 2003, Liu and Fang 2012], and a complex task with important economic potential [Kummamuru and Krishnapuram 2007].

For instance, Liu and Fang [2012] propose two methods to build entities profiles for research papers published in a specific track of a specific conference. In their work, Liu and Fang made an experiment profiling paper published in the Knowledge-Based Approaches (KBA) track of the 21st Text Retrieval Conference, TREC 2012. For this experiment, the authors consider 29 entities (topics) manually chosen from the English collection of Wikipedia that were representative of topics usually covered by KBA track papers along the previous editions.

Basically, Liu and Fang's methods perform the computation of a numerical score based on the number of occurrences of the entity names found in each paper. The methods differences rely on the use of weighting schemas to estimate the relevance of each occurrence according to the presence of co-occurrence of other entities. The conclusions of Liu and Fang indicate that these methods were effective to select relevant documents among the papers appearing in TREC 2012 proceedings.

Another related work worth mentioning is the paper authored by Xue and Zhou [2009] that proposes a method to perform text categorization using distributional features. This work does not explicitly mention the construction of entity profiles, but Xue and Zhou's method do create a descriptor of each possible category to be considered in the form of features. In such way, the category descriptors can be easily viewed as the category profile, and the categorization itself can be viewed as the computation of similarities between each category profile and each text features.

Putting our current work in perspective with these related works, our proposed technique carries on a profile building task that is similar to Xue and Zhou's category descriptors. The main difference of our approach, however, resides on the descriptors contents. While Xue and Zhou's techniques are generic features (number of words, *etc.*) found in the texts, our descriptors are remarkable terms (most relevant concept bearing terms) found in the texts. In this sense our work can be seen as an evolution of [De Souza et al. 2007].

Our proposed text categorization is similar to Liu and Fang's score computation, since we also compute a similarity index to estimate how related a text is to each entity. The main difference between Liu and Fang's and our approach resides in the specific score formulation. While Liu and Fang's observe co-occurrences of entities names, our approach weights more relevant concepts bearing terms found at the entities describing corpora and at the texts to categorize. In this sense, we revisit an old approach [Cavnar and Trenkle 1994], but we use a more effective term extraction.

3. Building Profiles Through Term Extraction from Corpora

The proposed technique starts creating entities descriptors, *i.e.*, a set of data associated to each entity that summarizes the relevant information for each entity. In our approach these descriptors are basically a set of relevant concept bearing terms found in the entity's defining corpus. To obtain these terms we perform a sophisticated term extraction procedure [Lopes and Vieira 2012] followed by a relevance index computation [Lopes et al. 2012]. Specifically, we submit the defining corpora of all entities to an extraction procedure that is actually performed in two steps: The texts are syntactically annotated by the parser PALAVRAS [Bick 2000]; The annotated texts are submitted to ExATOlp [Lopes et al. 2009] that performs the extraction procedure and relevance index computation. It is important to mention that our proposed technique can be applied with other tools to text annotation or term extraction with, at the authors best knowledge, no loss of generality.

Term extraction performed by ExATOlp delivers only concept bearing terms, since it only considers terms that are Noun Phrases (NP) and free of determiners (articles, pronouns, *etc.*). In fact, the extraction procedure performed by ExATOlp considers a set of linguistic based heuristics that delivers the state of the art concept extraction for Portuguese language texts [Lopes and Vieira 2012].

Term frequency, disjoint corpora frequency (*tf-dcf*) is also computed by ExATOlp. *tf-dcf* is an index that estimates the relevance of a term directly proportional to its frequency in the target corpus, and inversely proportional to its frequency in a set of contrasting corpora. Consequently, the computation of the relevance index requires not only the defining corpora, but also a set of contrasting corpora [Lopes et al. 2012].

Once the terms of the defining corpus for each entity are extracted and associated to their respective relevance indices, the proposed construction of each entity descriptor is composed by two lists of terms with their relevance indices:

- **top terms** - The first list is composed by the n top relevant terms[1], *i.e.*, the n terms with higher *tf-dcf* values;
- **drop terms** - The second list is composed by the n more frequent, but common, terms, *i.e.*, the terms with the higher frequency and lower *tf-dcf* values.

To rank the terms for the top terms list it suffices to rank the terms according to the *tf-dcf* index, which is numerically defined for term t in the target corpus c considering a set of contrasting corpora $\mathcal{G}$ as:

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

where $\textit{tf}_t^{(c)}$ is the term frequency of term t in corpus c.

To rank terms for the drop terms lists, it is possible to consider a relevance drop index numerically defined as the difference between the term frequency and the *tf-dcf* index, *i.e.*:

$$\textit{drop}_t^{(c)} = \textit{tf}_t^{(c)} - \textit{tf-dcf}_t^{(c)} \tag{2}$$

[1]The number of terms in each list is an arbitrary choice that is not fully analyzed yet. However, preliminary experiments indicate that lists of $n = 50$ terms seem effective.

An important point of the entity descriptors building process is to take into account the fact that sometimes distinct entities can have quite unbalanced corpora. This can be the result of entities with corpora with very different sizes, but it may also happen due to intrinsic characteristics of each defining corpus. In fact, even corpora with similar sizes can have very distinct occurrence distributions. Therefore, in order to equalize the eventual differences between values of distinct corpora we decided to adopt as numerical values of *tf-dcf* and *drop* indices not their raw value expressed by Eqs. 1 and 2, but the logarithm of those values. Such decision follows the basic idea formulated by the Zipf Law [Zipf 1935] that states that the distribution of term occurrences follows and exponential distribution. Consequently, adopting the logarithm values of *tf-dcf* and *drop* is likely to brings those index to a linear distribution[2].

Formally, the descriptor of each entity e, with $e \in \{1, 2, \ldots, E\}$, is denoted by the lists $\mathcal{T}_e$ and $\mathcal{D}_e$ composed by the information:

- $term(t_e^i)$ the i-th term of $\mathcal{T}_e$
- $idx(t_e^i)$ the logarithmic value of the *tf-dcf* of the i-th term of $\mathcal{T}_e$
- $term(d_e^i)$ the i-th term of $\mathcal{D}_e$
- $idx(d_e^i)$ the logarithmic value of the *drop* index of the i-th term of $\mathcal{D}_e$

Figure 1 describes this descriptor building process. In this figure, each entity is described by a defining corpus and from such corpus a term extraction and relevance index computation is made in order to generate a pair of lists to describe each entity.

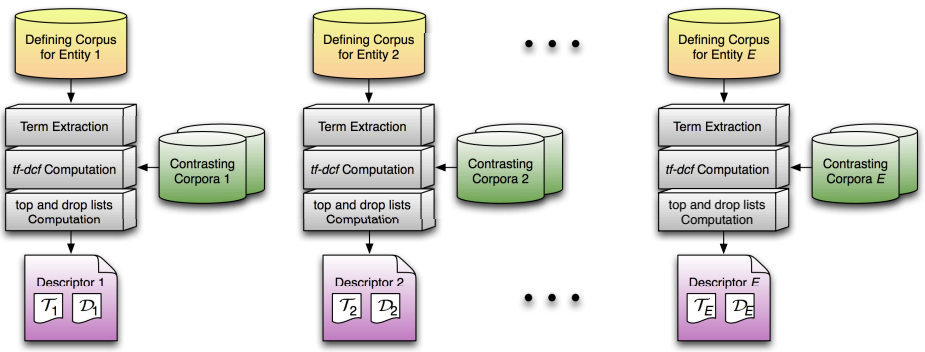

Figure 1. Descriptors Building Process

4. Applying Profiles to Categorize Texts

Given a set of entities dully characterized by their descriptors (top terms and drop terms lists), the categorization of a text (or corpus) can be made computing the similarity of such text (or corpus) with each entity. Obviously, the entity that is more similar to the text is considered the more adequate category.

[2]For the linearization purpose any logarithm would be enough. Specifically for this paper experiments a binary logarithm was adopted, but we also replicated the experiments with natural and decimal logarithms and, as expected, the overall results were not changed, *i.e.*, the numerical values of *tf-dcf* index changed, but the relevance ranking did not change.

Specifically, the proposed technique starts extracting the relevant terms for the text (or corpus) to categorize. This term extraction and relevance index computation must be made using the same tools and parameters as the ones used for constructing the entities descriptor, *i.e.*, in our case, the text to categorize must be submitted to PALAVRAS and ExATOlp with the same contrasting corpora. This step will produce a list of terms with their respective *tf-dcf* index. Analogously, to the profile indices, instead of the raw *tf-dcf* index, we will store its logarithm. Formally, such list is denoted $\mathcal{C}$ and it is composed by the information:

- $term(c^i)$ the i-th term of $\mathcal{C}$
- $idx(c^i)$ the logarithm of the *tf-dcf* index of the i-th term of $\mathcal{C}$

The similarity of a text to categorize with term list $\mathcal{C}$ to an entity e is computed by:

$$sim\ \mathcal{C}_e = \sum_{i=1}^{|\mathcal{C}|} idx(c^i) \left[top_e \left(term(c^i) \right) + drop_e \left(term(c^i) \right) \right] \tag{3}$$

where:

$$top_e \left(term(c^i) \right) = \begin{cases} idx(t_e^j) & \text{if} \quad term(c^i) = term(t_e^j) \\ 0 & \text{otherwise} \end{cases}$$

$$drop_e \left(term(c^i) \right) = \begin{cases} idx(d_e^j) & \text{if} \quad term(c^i) = term(d_e^j) \\ 0 & \text{otherwise} \end{cases}$$

Figure 2 describes this text (or corpora) categorization process. In this figure, the extracted terms of the text to categorize are compared to each entity descriptor, computing the similarity index for each entity.

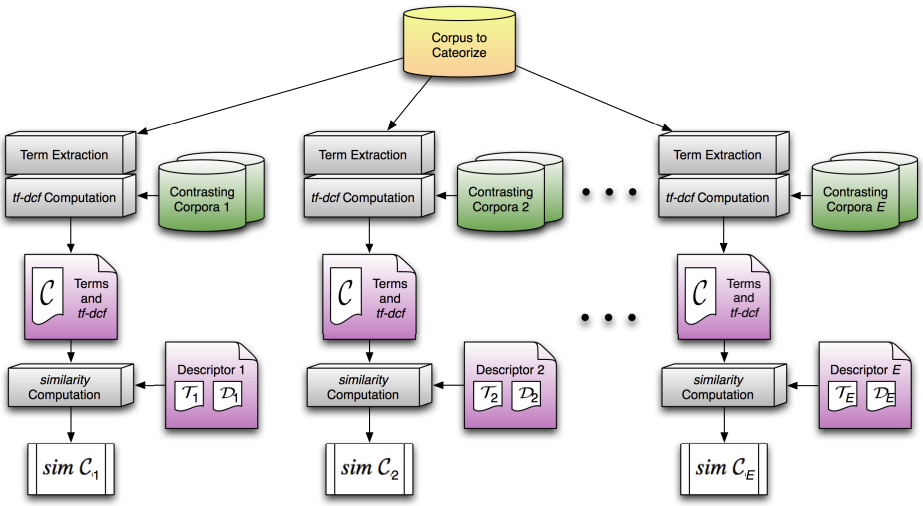

Figure 2. Corpus Categorization Process

5. Experiments for a Set of Professors

To illustrate the proposed technique, we conduct and experiment creating profiles for the full set of professors that successfully advised at least 5 M.Sc. thesis or Ph.D. dissertations from the creation of a Computer Science Graduate Program of a research intensive University from 1994 to 2013. In this corpora gathering process were kept only thesis and dissertation written in Portuguese to whom the text was electronically available. From a practical point of view, we managed to gather about 90% (370 of 410) of the published thesis and dissertations successfully presented during these 20 years. It resulted in 24 professors, grouped in 6 research groups. To each of these professors we assumed that their advised thesis and dissertations were their defining corpora.

Table 1 presents some information about these corpora. In this Table the name of professors was omitted and only a symbolic ID is presented. The name of the research groups is generically indicated by the acronyms BIO for Bioinformatics, AI for Artificial Intelligence, PD for Parallelism and Distribution, DES for Digital and Embedded Systems, SEDB for Software Engineering and Data Bases, and GHCI for Graphics and Human-Computer Interface. This division of research groups follows a classification based on current and historical groups of professors during this 20 years period. To each corpus this table also indicates the total numbers of texts, words and extracted terms.

Table 1. Entities and Corpora Characteristics

Professor	group	# texts	# words	# terms	Professor	group	# texts	# words	# terms
P01	BIO	9	187,010	39,859	P13	DES	19	506,457	108,958
P02	AI	6	101,331	21,722	P14	SEDB	21	635,691	139,911
P03	AI	13	219,930	44,707	P15	SEDB	20	441,555	92,986
P04	AI	25	587,177	120,772	P16	SEDB	28	512,899	103,491
P05	PD	16	287,923	60,727	P17	SEDB	16	425,069	87,532
P06	PD	22	391,329	89,575	P18	SEDB	11	290,040	62,774
P07	PD	14	310,905	64,193	P19	SEDB	5	120,199	24,051
P08	PD	15	278,346	59,582	P20	GHCI	13	223,323	48,089
P09	PD	25	431,082	90,501	P21	GHCI	12	285,893	62,432
P10	DES	8	164,740	34,267	P22	GHCI	11	203,938	42,065
P11	DES	12	269,171	59,297	P23	GHCI	13	197,942	43,534
P12	DES	24	591,018	122,594	P24	GHCI	12	164,130	32,544

5.1. Building Descriptors

To build the descriptors for the 24 entities according to the process described in Section 3, we consider the following:

- All thesis and dissertation advised were assumed to be the adequate description of each professor research topics, and, therefore, all texts advised by a professor were considered his/her defining corpus;
- For *tf-dcf* relevance index computation, the texts of all research groups, but the one to whom the professor belongs, were considered as contrasting corpora;
- The top terms and drop terms lists were limited to 50 terms and their respective indices (*tf-dcf* and *drop*).

Finally, the aimed 24 entities descriptors were composed by 24 pairs of lists (a pair for each professor) denoted $\mathcal{T}_e$ and $\mathcal{D}_e$, with $e \in \{\text{P01, P02, } \ldots, \text{P24}\}$.

5.2. Categorization of Texts

To illustrate the effectiveness of the builded entity profiles to categorize texts (or corpora) we conduct six experiments:

1. We took a conference paper written by one professor from PD research group (5 thousand words);
2. We took a short note on the Bioinformatics domain (1 thousand words);
3. We took a M.Sc. thesis on NLP - Natural Language Processing absent from the defining corpora (13.6 thousand words);
4. We took a corpus on DM - Data Mining with 53 texts (1.1 million words);
5. We took a corpus on SM - Stochastic Modeling with 88 texts (1.1 million words);
6. We took a corpus on Pneumology with 23 texts (16.5 thousands of words).

In all experiments, we perform the proposed process (Section 4) to extract terms using the same contrasting corpora. Consequently, each text (or corpus) was submitted to 6 different sets of contrasting corpora, *e.g.*, when computing similarity for a professor from research group PD, the contrasting corpora were the texts from all professors from other research groups (BIO, AI, DES, SEDB and GHCI). Table 2 presents the top ten entities (e), *i.e.*, group and professor id., according to the computed similarity (*sim* $\mathcal{C}_e$).

Table 2. Top Ten Entities According to Computed Similarity

Exp. 1 - PD		Exp. 2 - BIO		Exp. 3 - NLP	
e	*sim* $\mathcal{C}_e$	e	*sim* $\mathcal{C}_e$	e	*sim* $\mathcal{C}_e$
PD - **P06**	5.33	BIO - **P01**	22.04	AI - **P03**	61.27
PD - P08	2.57	GHCI - P21	0.01	AI - **P04**	48.99
DES - P12	0.63	SEDB - P16	0.01	AI - P02	12.13
DES - P13	0.48	SEDB - P15	0.00	DES - P11	1.68
DES - P11	0.46	GHCI - P22	0.00	GHCI - P20	0.34
PD - P05	0.45	SEDB - P14	0.00	GHCI - P22	0.10
DES - P10	0.05	SEDB - P18	0.00	DES - P10	0.08
PD - P07	0.03	GHCI - P23	0.00	SEDB - P15	0.07
SEDB - P15	0.02	PD - P05	0.00	BIO - P01	0.06
SEDB - P17	0.02	AI - P04	0.00	SEDB - P18	0.06

Exp. 4 - DM		Exp. 5 - SM		Exp. 6 - Pneumo	
e	*sim* $\mathcal{C}_e$	e	*sim* $\mathcal{C}_e$	e	*sim* $\mathcal{C}_e$
SEDB - **P17**	422.8	PD - **P09**	1,737	BIO - **P01**	8.71
AI - P04	132.4	DES - P12	176	GHCI - **P23**	8.69
SEDB - P16	124.9	PD - P07	118	GHCI - **P22**	5.70
GHCI - P23	66.4	PD - P08	110	GHCI - **P20**	3.71
AI - P03	59.6	DES - P13	97	PD - P06	2.09
GHCI - P20	54.7	PD - P05	71	GHCI - P21	0.52
GHCI - P24	46.3	AI - P02	60	DES - P13	0.45
BIO - P01	44.4	BIO - P01	57	SEDB - P14	0.42
SEDB - P18	31.0	DES - P10	54	GHCI - P24	0.10
GHCI - P22	30.8	AI - P04	51	SEDB - P17	0.08

The first experiment (a conference paper written by P06) was clearly categorized for this professor. It is also remarkable that other professors from PD and DES research groups were also well ranked by the similarity.

The second experiment (a short note about Bioinformatics) was also a clear case to categorize, since it was clearly situated in the professors P01 expertise. Since P01 is the only researcher of BIO group, the results indicate clearly this entity as the more similar one.

The third experiment (a M.Sc. thesis on NLP) is also a clear categorization result, since the three top ranked professors were from AI research group, which comprises the area of NLP. It is also noticeable that professors P03 and P04 clearly dominated the similarity measure with a numerical value above and around 50, while the similarity for the others professors are around or less than 10. It is not a coincidence that these two professors concentrate their research on NLP.

The fourth experiment (DM corpus) is also an interesting result, since it clearly indicates a predominance of P17 that works on the subject of Data Warehouses. The two next top ranked professors are from SEDB and AI. Such result also makes sense, since many Data Mining techniques are strongly related to both Data Bases and Artificial Intelligence.

The fifth experiment (SM corpus) looks like the clearest result, since P09 main research is on the development of performance models and its similarity value (over 1,700) is much higher than the values for all other professors (less than 200). Accentuate the success of this experiment the observation that professors from PD and DES groups clearly dominate the highest similarity values.

The sixth experiment (Pneumology corpus) was chosen to illustrate how a topic far from the professors expertise would be categorized. None of the professors works on the topic of Pneumology, therefore, we would expect that none of the similarity values would clearly stand out from the others. Nevertheless, to our surprise some professors on subjects that could be related to the medical topics delivered the top four similarity values. This is likely to be an effect of some common terms found in Bioinformatics (P01) and also in human related topics (P23, P22 and P20).

Table 3. Ratio Between the Highest Similarity and Logarithm of Number of Words

	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6
highest *sim* $\mathcal{C}_e$	21.40	22.04	53.99	287.67	716.24	10.34
log_2 # words	12.29	9.97	13.73	20.07	20.07	14.01
ratio	1.05E-03	2.21E-00	4.51E-03	3.84E-04	1.58E-03	5.28E-04

Finally, a clear observation from the results in Table 2 is the quite distinct values obtained for each experiment. We noticed a clear, and expected, relation between the size of the texts to categorize and the numerical values of the similarity. In Table 3 we observe the ratio between the highest similarity value and the binary log of the number of words in the texts for each experiment. This ratio seems to indicate the level of confidence in the categorization, *e.g.*, for Experiments 4 and 6 the confidence is lower than the others. On the contrary, Experiment 2 outcome seems to be very reliable, and not Experiment 5 as it would appear in the first observation.

6. Conclusion

This paper proposed a technique to build entity profiles according to a guided term extraction taking relevance indices into account. The builded profiles were applied to a categorization task with a considerable success as shown in the six presented experiments. Therefore, this paper contribution is two-fold, since both entity profiles building and text categorization are interesting problems tackled by the proposed technique.

The entity profiles building process based on term extraction producing top terms and drop terms lists is a robust and innovative solution to a complex problem that can potentially solve many practical issues. Besides text categorization, other possible applications are automatic authoring recognition; terminology classification; *etc.*

The text categorization process based on the entities profiles is a direct application with many practical uses. For instance, the conducted experiments over the M.Sc. thesis and Ph.D. dissertations of a graduate program can be very useful to help practical decisions like: which candidate is more adequate to a future advisor; which professor is the best placed to evaluate an external project or publication; which professors are the more adequate to compose a jury; *etc.* Nevertheless, it is important to keep in mind that our main goal is to propose a profiling technique and the text categorization was just an application example.

Our experiments are the first tests of this original profiling technique, and natural future work for our research will be the deep analysis of parameters as the size of descriptor lists (n), impact of a very large number of entities, *etc.* It is also a possible future work the broader experimentation over other data sets, and even other applications than text categorization. Anyway, the presented results are encouraging due to the effectiveness achieved, specially for large amounts of text to categorize.

References

Balog, K., Ramampiaro, H., Takhirov, N., and Nørvåg, K. (2013). Multi-step classification approaches to cumulative citation recommendation. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 121–128, Paris, France, France. Le Centre des Hautes Etudes Internationales d'Informatique Documentaire.

Bick, E. (2000). *The parsing system PALAVRAS: automatic grammatical analysis of portuguese in constraint grammar framework*. PhD thesis, Arhus University.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

De Souza, A., Pedroni, F., Oliveira, E., Ciarelli, P., Henrique, W., and Veronese, L. (2007). Automated free text classification of economic activities using vg-ram weightless neural networks. In *Intelligent Systems Design and Applications, 2007. ISDA 2007. Seventh International Conference on*, pages 782–787.

Kummamuru, K. and Krishnapuram, R. (2007). Method, system and computer program product for profiling entities. US Patent 7,219,105.

Liu, X. and Fang, H. (2012). Entity Profile based Approach in Automatic Knowledge Finding. In *Proceedings of Text Retrieval Conference, TREC 2012*.

Lopes, L., Fernandes, P., and Vieira, R. (2012). Domain term relevance through tf-dcf. In *Proceedings of the 2012 International Conference on Artificial Intelligence (ICAI 2012)*, pages 1001–1007, Las Vegas, USA. CSREA Press.

Lopes, L., Fernandes, P., Vieira, R., and Fedrizzi, G. (2009). ExATOlp – 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)*, pages 427–431, Poznan, Poland. Faculty of Mathematics and Computer Science of Adam Mickiewicz University.

Lopes, L. and Vieira, R. (2012). Heuristics to improve ontology term extraction. In *PROPOR 2012 – International Conference on Computational Processing of Portuguese Language*, LNCS vol. 7243, pages 85–92.

Wei, L. (2003). Entity Profile Extraction from Large Corpora. In *Proceedings Pacific Association of Computational Linguistics 2003*.

Zhou, M. and Chang, K. C.-C. (2013). Entity-centric document filtering: Boosting feature mapping through meta-features. In *Proceedings of the 22Nd ACM International Conference on Conference on Information; Knowledge Management*, CIKM '13, pages 119–128, New York, NY, USA. ACM.

Zipf, G. K. (1935). *The Psycho-Biology of Language - An Introduction to Dynamic Philology*. Houghton-Mifflin Company, Boston, USA.