Stopwords Identification by Means of Characteristic and Discriminant Analysis

Giuliano Armano, Francesca Fanni and Alessandro Giuliani
Department of Electrical and Electronic Engineering (DIEE), University of Cagliari, via Marengo 2, 09123 Cagliari, Italy

Keywords: Stopwords, Discriminant Capability, Characteristic Capability, Text Classification.

Abstract: Stopwords are meaningless, non-significant terms that frequently occur in a document. They should be removed, like a noise. Traditionally, two different approaches of building a stoplist have been used: the former considers the most frequent terms looking at a language (e.g., english stoplist), the other includes the most occurring terms in a document collection. In several tasks, e.g., text classification and clustering, documents are typically grouped into categories. We propose a novel approach aimed at automatically identifying specific stopwords for each category. The proposal relies on two unbiased metrics that allow to analyze the informative content of each term; one measures the discriminant capability and the latter measures the characteristic capability. For each term, the former is expected to be high in accordance with the ability to distinguish a category against others, whereas the latter is expected to be high according to how the term is frequent and common over all categories. A preliminary study and experiments have been performed, pointing out our insight. Results confirm that, for each domain, the metrics easily identify specific stoplist wich include classical and category-dependent stopwords.

1 INTRODUCTION

Stopwords are meaningless terms that frequently occur in a document. They usually have no real purpose in describing document contents and they cannot discriminate between relevant and non-relevant items. Hence, such terms should be removed from documents before processing with Machine Learning (ML) or Information Retrieval (IR) procedures. Indeed, non-significant terms represent noise and, dependly on the adopted techinique, may actually reduce retrieval effectiveness. Classical lists of stopwords (*stoplists* hereinafter) include only the most frequently occurring terms in a specific language. However, removing very frequent terms could also reduce performances. With the goal of maximizing the performance of a ML or IR task, it is useful to devise methods and algorithms able to automatically and dynamically build different stoplists, depending on the given dataset collection. State-of-the-art algorithms for the automatic identification of stopwords currently rely on the entire document collection for building a unique stoplist (Lo et al., 2005; Sinka and Corne, 2003b,a). In several tasks, e.g., text categorization, data items are typically grouped into categories. We assume that identifying a proper and dynamic stoplist for each category should improve the performance of an IR task. For example, let us consider a dataset for the domain "Sport", containing the categories "Volleyball", "Basket", and "Football"; intuitively, for the given domain the term *ball* should be considered a stopword, as it is not relevant to discriminate among the cited categories of "Sports".

In this paper, we perform a preliminary study on the behavior of stopwords in document collections, and we propose a stoplist detection approach. The work is based on novel metrics able to assess the discriminant and characteristic capability. For each term, the former is expected to be high in accordance with the ability to distinguish a given category against others. The latter is expected to be high according to how the term is frequent and common over all categories.

The rest of the paper is structured as follows: Section 2 reports the background and the related work about automatic stopwords identification; Section 3 describes the adopted metrics; in Section 4 the behavior of stopwords in the space defined by the metrics is shown. Experiments are reported in Section 5; Section 6 discusses the strengths and the weaknesses of this work. Section 7 ends the paper with conclusions and future work.

Armano G., Fanni F. and Giuliani A..
Stopwords Identification by Means of Characteristic and Discriminant Analysis.
DOI: 10.5220/0005194303530360
In *Proceedings of the International Conference on Agents and Artificial Intelligence* (ICAART-2015), pages 353-360
ISBN: 978-989-758-074-1

2 BACKGROUND

In this work, the underlying scenario is text categorization, where source items are textual documents (e.g., webpages, online news, scientific papers, and e-books).

According to Luhn (1958), in a document a relatively small number of terms are meaningful for a ML or IR task. Non-informative terms that frequently occur in a document are called stopwords. Such terms are mainly pronouns, articles, prepositions, conjunctions, frequent verbs forms, etc. (Silva and Ribeiro, 2003). In principle, stopwords are expected to occur in every document. The work of Francis and Kucera (1983) show that the ten most frequent terms in the English language typically occur between 20 and 30 percent of the total number of terms in a document collection. Furthermore, Hart (1994) assesses that over 50% of all terms in an English document belongs to a set of about 135 common terms in the Brown corpus (Kucera and Francis, 1967).

Stopwords are expected not only to have a very low discriminant value, but often they could introduce noise for an IR task (Rijsbergen, 1979). For these reasons, a stoplist is usually built with terms that should be filtered in the document representation process, since they actually reduce retrieval effectiveness. Traditionally, stoplists are supposed to have included only the most frequently occurring terms in a specific language. Several systems have been developed for suggest stoplists in an automatic manner. SMART (Salton, 1971) has been the first system that automatically built a stoplist, containing 571 English terms. Fox (1989) initially proposed only 421 terms, and then derived a stoplist from the Brown corpus (Francis and Kucera, 1983). This set was typically adopted as standard stoplist in many subsequent research works and systems (Fox, 1992).

Nonetheless, the use of fixed stoplists across different document collections could negatively affect the performance of a system. In English, for example, a text classifier might encounter problems with terms such as “language c”, “vitamin a”, “IT engineer”, or “US citizen” where the forms “c”, “a”, “it”, or “us” are usually removed (Dolamic and Savoy, 2010; Lo et al., 2005). In other words, we deem that each document collection is unique, making useful to devise methods and algorithms able to automatically build different stoplist for each collection, with the goal of maximizing the performance of a ML or IR system.

Several metrics are used to weight terms for identifying a stoplist in a document collection. The most common metric is the TF-IDF (Salton and McGill, 1984), in which the weight is given as a product of two parts: the *term frequency* (TF), i.e., the frequency of a term in a document; and the *inverse document frequency* (IDF), i.e., the inverse of the number of documents in the collection in which the term occurs. The use of TF-IDF makes possible to rank terms, filtering whose that frequently appear in a document collection (Silva and Ribeiro, 2003). A further approach to find stopwords is the use of entropy as discriminant measure (Sinka and Corne, 2003b). Entropy, here, is correlated with the frequency variance of a given term over multiple documents, meaning that terms with very high frequency in some documents, but very low frequency in others, will have higher entropy than terms with similar frequency in all documents of the collection. The list of terms is ordered by ascending entropy to reveal terms that have a greater probability of being noisy. Further works define automated stopwords extraction techniques by focusing on statistical approaches (Hao and Hao, 2008; Wilbur and Sirotkin, 1992).

As the most acknowledged approaches do not give a value to the discriminant power of a term, we use novel metrics able to measure it, with the goal of identifying stopwords for a document collection. The adopted metrics are the discriminant and the characteristic capability defined in a previous work (Armano, 2014). The former is expected to raise in accordance with the ability to distinguish a given category against others. The latter is expected to grow according to how the term is frequent and common over all categories. In our work, terms having a low discriminant value and high characteristic value are considered stopwords.

3 THE ADOPTED METRICS

In this paper, we adopt two novel metrics able to provide relevant information to researchers in several IR and ML tasks (Armano, 2014). The proposal consists in two unbiased metrics, i.e., independent from the imbalance between *positive* (P) and *negative* (N) samples. For a binary classifier, the former means that the item belongs to the considered category C, whereas the latter means that the item belongs to the alternate category $\overline{C}$ (i.e., the set of remaining categories). The metrics rely on the classical four basic components of a confusion matrix, i.e., true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The most acknowledged metrics, e.g., *precision*, *recall* (or *sensitivity*), and *accuracy* are calculated in terms of such entries. Similarly, our metrics relies on some of the classical metrics (see Table 1 for

Table 1: The adopted classical metrics.

Sensitivity	$\rho = \dfrac{TP}{TP+FN} = \dfrac{TN}{P}$
Specificity	$\bar{\rho} = \dfrac{TN}{TN+FP} = \dfrac{TN}{N}$
Fall-out	$(1-\bar{\rho}) = \dfrac{FP}{TN+FP} = \dfrac{FP}{N}$

their definitions[1]).

In this work, we adopt a pair of unbiased metrics able to capture the concepts of discriminant and characteristic capabilities. The underlying semantics is the straightforward: a metric devised to measure the former capability is expected to raise in accordance with the ability of separating positive from negative samples, whereas the latter is expected to raise in accordance with the ability of aggregating positive and negative samples. To our knowledge, no previous works provide satisfactory definitions able to account for the need of capturing the potential of a model according to its discriminant and characteristic capability. The definitions of *discriminant* (δ) and *characteristic* (φ) metrics are the following:

$$\delta = Sensitivity - Fall\ out \tag{1}$$

$$\varphi = Sensitivity - Specificity \tag{2}$$

Replacing sensitivity, specificity, and fall-out with their formulas, we obtain:

$$\delta = \rho + \bar{\rho} - 1 = \frac{TP}{P} - \frac{FP}{N} \tag{3}$$

$$\varphi = \rho - \bar{\rho} = \frac{TP}{P} - \frac{TN}{N} \tag{4}$$

The above metrics show the following behavior: δ is high when a classifier partitions a given set of samples in accordance with the corresponding class labels; on the other hand, φ is high when a classifier tends to cluster negative and positive samples together. Let us note that here the conceptualization of "characteristic property" affects all samples, regardless from the corresponding class label, whereas the classical definition adopted in ML focuses only on samples that belong to the main class (Armano, 2014).

Assuming both ranging from -1 to +1, the proposed metrics show an orthogonal behavior, as summarized in the Table 2.

It has been proved that the $\varphi - \delta$ space is constrained by a rhomboidal shape (Armano, 2014), as reported in Figure 1.

Table 2: The behavior of the proposed metrics.

Name	**Domain**	**Expected Behavior**
δ	[-1, +1]	$\delta \cong \pm 1 \Leftrightarrow \varphi \cong 0$
φ	[-1, +1]	$\varphi \cong \pm 1 \Leftrightarrow \delta \cong 0$

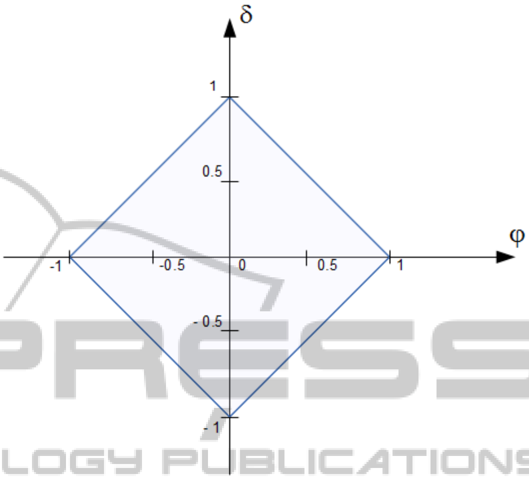


Figure 1: The theoretical $\varphi - \delta$ space.

In this work we apply the metrics in a text categorization context, meaning that, for each term of the given domain, they are able to evaluate its discriminant and characteristic capability. For each term, the former is expected to grow in accordance with the ability to distinguish a given category against others. The latter is expected to grow according to how the term is frequent and common over all categories. In order to give a suitable definition of φ and δ in this scenario, we firstly introduce the confusion matrix entries for the considered context. In text classification, a generic term t contained in a document represents the sample under analysis, and it can be considered as *positive* if the document containing t belongs to the main class C; on the other hand, t is considered as *negative* if the document containing it belongs to the alternate class $\overline{C}$ (see Table 3, in which the absence of term is denoted as $\bar{t}$).

Table 3: Confusion matrix entries in text classification.

Semantics	Denoted as	Description
TP	$\#(t,C)$	#docs of C containing t
FP	$\#(t,\overline{C})$	#docs of $\overline{C}$ containing t
FN	$\#(\bar{t},C)$	#docs of C NOT containing t
TN	$\#(\bar{t},\overline{C})$	#docs of $\overline{C}$ NOT containing t
P	$\#(C)$	#docs of C
N	$\#(\overline{C})$	#docs of $\overline{C}$

[1] Note that $TP+FN=P$ and $TN+FP=N$

In so doing, we can define φ and δ in this scenario as reported in formulas 5 and 6.

$$\delta = \frac{\#(t,C)}{\#(C)} - \frac{\#(t,\overline{C})}{\#(C)} \tag{5}$$

$$\varphi = \frac{\#(t,C)}{\#(C)} - \frac{\#(\overline{t},\overline{C})}{\#(C)} \tag{6}$$

4 STOPWORD IDENTIFICATION

We expect that important terms for text classification appear in upper and lower corner of the rhombus in Figure 1, as they have high values of |δ|. In particular, a high positive value of δ means that the term is highly discriminant for identifying positive samples, whereas a high negative value of δ means that the term is highly discriminant for identifying negative documents. As for the characteristic capability, terms that occur barely on documents are expected to appear in the left corner of the rhombus (high negative values of φ), while stopwords are expected to appear in the right handed corner (high positive value of φ). Figure 2 outlines the described behavior for all cases.

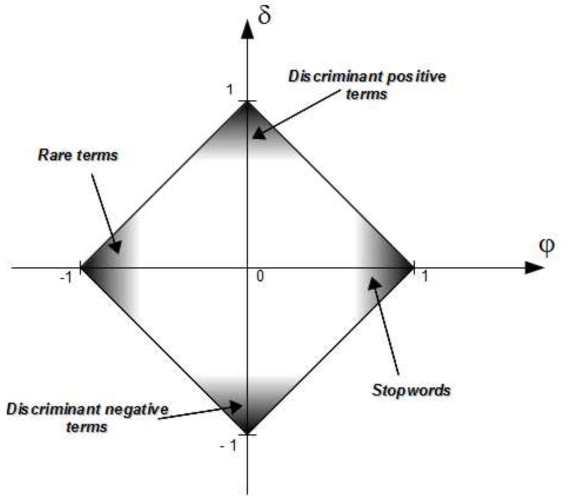


Figure 2: The role of the terms.

It is worth pointing out that terms falling in the right handed corner do not necessarily represent typical stopwords *only* (i.e., common articles, nouns, conjunctions, verbs, and adverbs). Rather, also category-dependent stopwords may be located in that area.

5 EXPERIMENTS

We use a collection of webpages, extracted from the DMOZ taxonomy[2]. DMOZ is the collection of HTML documents referenced in a Web directory developed in the Open Directory Project (ODP). We selected 174 categories organized in 36 domains, with a total of about 35000 documents. Each domain consists in a set of siblings nodes. The goal is to automatically build stoplists for specific domains, e.g., stoplist for the domain “space”, or for the domain “music”. First the given set of documents must be converted to a suitable representation. The most common approach is to tokenize strings of characters, in order to obtain a bag of words. Hence, textual information from each page is extracted, and noisy elements (e.g., tags and meta-data) removed. A document is then represented as bag of words, each term being weighted with two values: the discriminant and the characteristic capabilities, computed by applying equations 5 and 6.

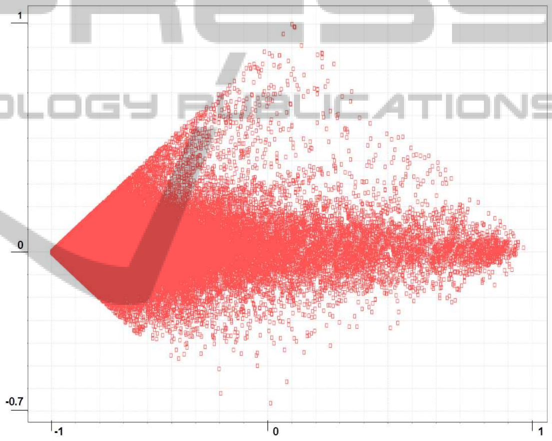


Figure 3: Rhombus obtained plotting the characteristic and the discriminant value of each term in all domains.

Figure 3 depicts the distribution of the data points in terms of discriminant and characteristic capabilities in all domains. As expected, the points fall in the rhombus area as pointed out in Table 2 and displayed in Figure 1. Terms are concentrated on the left corner of the rhombus, meaning that most terms tend to be rare and uncommon in the dataset. This is in accordance with to the Zipf’s law (Zipf, 1935), that proves that, in a given corpus, the frequency of any term is inversely proportional to its rank in the frequency table. Ideally, the most frequent term will occur approximately twice as often as the second most frequent term, three times as often as the third most frequent term, etc. In Figure 4, a log-log chart is depicted, in which each point is related to a term; the *x* axis reports the rank of a term, and the *y* axis reports its number of occurrences. The points follow the Zipf’s Law (the

[2] http://www.dmoz.org

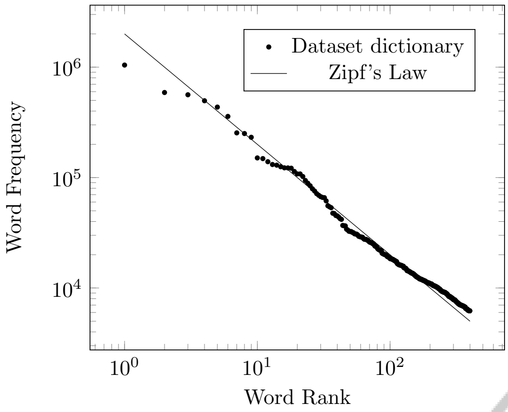


Figure 4: Zipf's Law for the dataset vocabulary.

straight line), as pointed out in the chart.

Our proposal is based on the insight that also category-dependent stopwords fall in the right handed corner. To further investigate this issue, we selected several domains from the dataset, each containing up to six categories. For each category we performed experiments aimed at projecting terms in the $\varphi - \delta$ space, to verify whether the expected behavior is confirmed. Being interested in the right handed corner of the $\varphi - \delta$ space, we decided to take into account only terms with a characteristic value greater than 0.

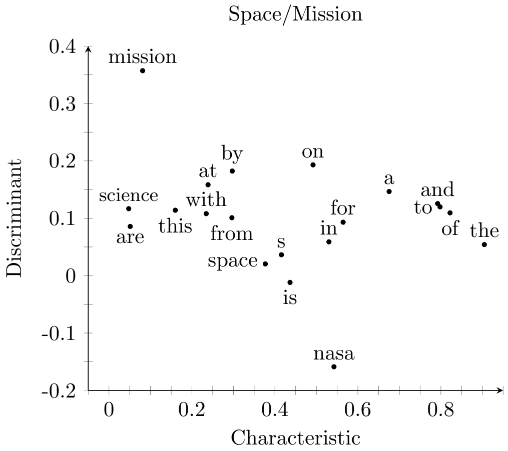


Figure 5: Words with $\varphi > 0$ for category "Mission".

An example is reported in Figure 5, concerning the "Mission" category (belonging to the domain "Space"). Each term having $\varphi > 0$ is reported in the $\varphi - \delta$ diagram. As expected, most terms belong to a classical stoplist. However, several terms are apparently related to the given category (i.e., the terms *mission*, *space*, *science*, and *nasa*). Although, for the sake of brevity, we have reported here only this example, the co-occurrence of global and category-specific terms is confirmed in all categories of the dataset.

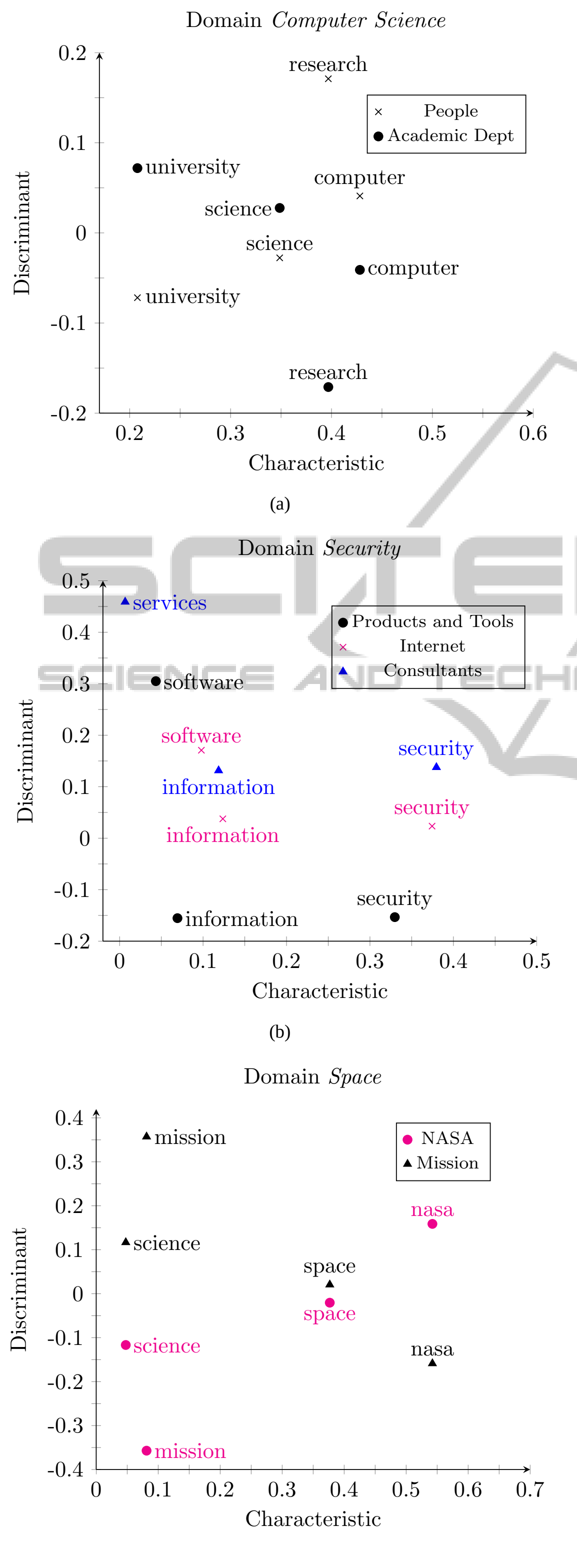
Subsequently, we filtered out the classical stopwords, and analyzed category-specific terms only. Three different domains of the DMOZ taxonomy (i.e., *Computer Science*, *Security*, and *Space*) are investigated. *Computer Science* contains the categories *Academic Departments* and *People*; *Security* contains the categories *Products and Tools*, *Internet*, and *Consultants*, whereas *Space* contains the categories *NASA* and *Mission*.

As described in the previous Sections, a low $|\delta|$ and a high φ identify a stopword, while terms with a high value of $|\delta|$ are meaningful for the specific category. Hence, recalling that global stopwords have already been removed, Figure 6 highlights that not all the remaining terms are specific stopwords. In particular, terms like *computer* or *science*, intuitively, seem to be common in the categories belonging to *Computer Science* domain. This aspect is confirmed in the Figure 6(a), in which such terms have low δ, and high φ. In the same way, the term *security* is obviously common in the categories belonging to the domain *Security*. The term, intuitively, could be discriminant in a more general domain containing the category *Security*.

Conversely, in Figure 6(b) the term *services* has high discriminant capability for the category *Consultants*. Similarly, in Figure 6(c) the term *mission* has high discriminant capability (with respect to the characteristic value) for the category named with the same term (Mission). Reasonably, the term is clearly representative of the category *Mission*. Furthermore, the term *mission* is highly "negative" discriminant for the alternate category (in this case it is only the category *NASA*), meaning that it represents a negative sample for the category *NASA*.

Here, the focus is to filter out terms having significantly high discriminant value, and considering only terms with a reasonable value of characteristic. The filtered terms are considered stopwords. A simple filtering criteria is to consider terms having $\varphi_{term} > |\delta_{term}|$, and the set of terms that respect this constraint represents the stoplist of a given category. Looking at Figure 6(a), we can assert that every term respects the constraint described above. Furthermore, for the domain *Computer Science*, the category-dependent stoplists of *People* and *Academic Departments* contain the same stopwords. Hence, the set of category-dependent stopwords for this domain is easily represented by terms *computer*, *research*, *science*, and *university*. These stopwords should be included in a classical stoplist for defining the total stoplist of the domain.

In Figure 4, the terms that not respecting the filtering criteria are discarded. The remaining terms,

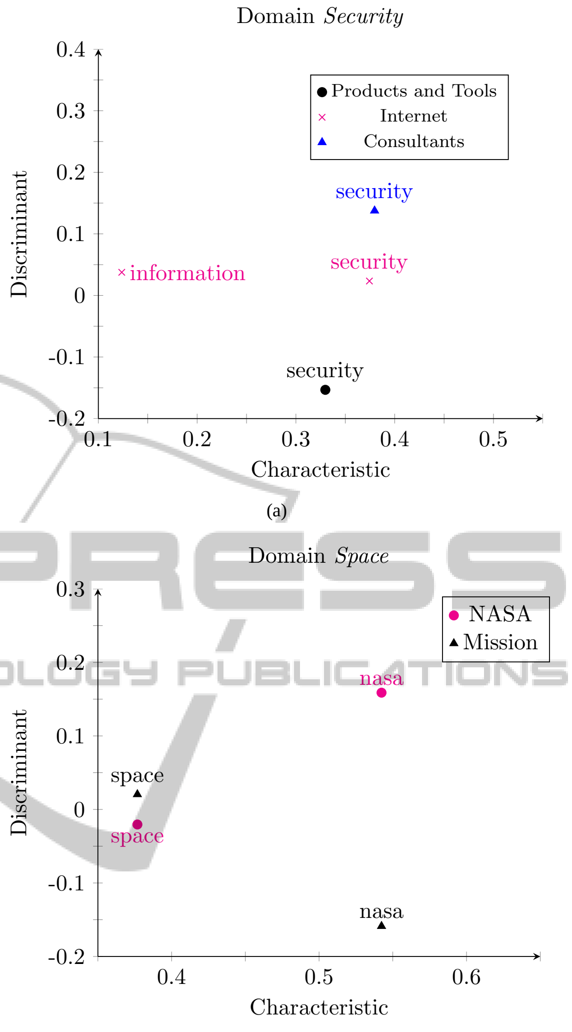


(a)

(b)

(c)

Figure 6: Non-classical terms for the domains *Computer Science*, *Security*, and *Space*.



(a)

(b)

Figure 7: Category-dependent stopwords for domains *Security* and *Space*.

in our view, compose the set of category-dependent stopwords.

Figure 4(a) points out that there are a common category-dependent stopword (*security*), and one stopword for the sole category *Internet* (the term *information*). This is due to the fact that, for *Product and Tools* and *Consultants*, the constraint $|\varphi_{term}| > |\delta_{term}|$ is not respected by a small difference between φ_{term} and $|\delta_{term}|$, as showed in Figure 6(b). Furthermore, the term has low φ and low δ in each category (around 0.1). We focus on this issue in Section 6.

Figure 4(b) reports the category-dependent stopwords for the categories rooted by *Space* (*Mission* and *Nasa*). Both categories contains the terms *space* and *nasa*. Note that here, the term *nasa* should be discriminant for the *Nasa* category. Nonetheless, intuitively, a webpage categorized in DMOZ with the class *Space* has a high probability of containing the term *nasa*, as most space missions are related to the NASA organization, and it is not surprising that the term is considered a stopword.

Finally, we could define a set of stoplists for the three domain under analysis, integrating the classical stopwords with the category-dependent ones. Table 4 reports the stoplists, in which each list is defined by considering global and category-dependent stopwords[3].

Table 4: Stoplists for specific domains.

Computer Science	Security	Space
of and ... computer ... science ... research ... university	and the to a ... security ... information ...	the of to and ... nasa ... space ...

6 DISCUSSION

The experiments show the potential usefulness of the adopted metrics in automatically defining domain-specific stoplists. As the work is in a preliminary stage, some issues are currently under study. In our opinion, the proposal encourages further studies on improving the approach. As reported in the previous Section, the approach allows to easily and quickly identify stoplists for any given domain. The adopted metrics are unbiased, that is, they are independent on the imbalance between positive and negative samples. We deem that the stoplists obtained with the proposal are useful as feature selection method. A future stage of the work is the experimentation of this method in text classification tasks.

The experiments remark a further behavior of category-dependent stopwords, i.e., they tend to be centered in the middle of the positive side of φ (see Figure 8). In our view, this phenomenon is caused by two main issues: (i) due to the nature of the dataset, although category-dependent stopwords appear in most documents, they have a distribution for which the probability to appear in almost the totality of documents is not enough high; (ii) the categories could contain distinct clusters of documents (i.e., further sub-categories); this is reasonable, if we take into account that the leaves of our dataset are extracted from a deeper taxonomy (DMOZ), and they are built as the union of their children.

Figure 8 gives a graphical representation of the phenomenon .

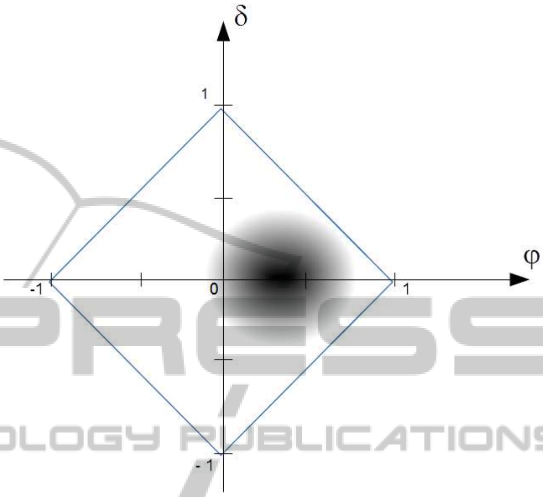


Figure 8: The disposition of category-dependent stopwords.

Currently, we are focusing on the behavior of terms falling in a neighborhood of the origin. Theoretically, if a term has a zero value for both δ and φ in a given category *C*, it is equally distributed in the domain in this way: half of *C* documents contain the term, and also half of documents of the alternate category $\overline{C}$ contain the term. If a term is projected close to the origin of the space, there is uncertainty in considering the term as stopword or not. There is the need of defining a suitable filtering criteria, able to select the most appropriate category-dependent stopwords. The main idea is to define, in the $\varphi - \delta$ space, the area of stopwords by devising suitable functionals, able to capture the real nature of terms falling in the uncertainty area. An example of functional is expected to be a constant value, for which we could rewrite the criteria as $\varphi_{term} - |\delta_{term}| > \varepsilon$, and let ε varying in order to study the behavior.

7 CONCLUSIONS AND FUTURE WORK

Stopwords are meaningless terms that frequently occur in a document. In this work, a novel approach for the automatic identification of stopwords has been

[3] For the sake of visualization, we do not report every global stopwords in the list

proposed. A stoplist should contain not classical stopwords only; rather, it should include terms that can be frequent, common, and non informative for identifying the category of a textual document.

The proposed approach is based on two novel metrics able to measure the discriminant and the characteristic capabilities (respectively, φ and δ) of a term in a specific domain. Indeed, the adopted metrics permit to identify also category-dependent stopwords, useful for reducing noise and improving the performance of an IR or ML task. Such terms are dependent on the considered domain. Our proposal is to consider as stopwords all terms having a high φ and low δ. A preliminary study and experimentations have been performed, pointing out our insight. Results confirmed that the stopwords, classical and category-dependent, tend to confirm the theoretical behavior, placing in the right handed corner of a rhombus area in the φ-δ space. As for future work, we are currently studying the most appropriate thresholds of φ and δ for selecting the stopwords. Furthermore, we are planning to perform experiments in text classification tasks, in order to evaluate the usefulness of the dynamic identification of stopwords. In fact, we suppose that the metrics can be used as a feature selection approach. Finally, we are setting up further experiments on different datasets.

ACKNOWLEDGEMENTS

This work has been supported by LR7 2009 - Investment funds for basic research (funded by the local government of Sardinia).

REFERENCES

Armano, G. (2014). A direct measure of discriminant and characteristic capability for classifier building and assessment. Technical report, DIEE, Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy. DIEE Technical Report Series.

Dolamic, L. and Savoy, J. (2010). When stopword lists make the difference. *J. Am. Soc. Inf. Sci. Technol.*, 61(1):200–203.

Fox, C. (1989). A stop list for general text. *SIGIR Forum*, 24(1-2):19–21.

Fox, C. (1992). Information retrieval. chapter Lexical Analysis and Stoplists, pages 102–130. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Francis, W. N. and Kucera, H. (1983). *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin.

Hao, L. and Hao, L. (2008). Automatic identification of stop words in chinese text classification. In *CSSE (1)'08*, pages 718–722.

Hart, G. W. (1994). To decode short cryptograms. *Commun. ACM*, 37(9):102–108.

Kucera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence, RI.

Lo, R. T.-W., He, B., and Ounis, I. (2005). Automatically building a stopword list for an information retrieval system. *JDIM*, 3(1):3–8.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165.

Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition.

Salton, G. (1971). *The SMART Retrieval System-Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Salton, G. and McGill, M. (1984). *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.

Silva, C. and Ribeiro, B. (2003). The importance of stop word removal on recall values in text categorization. In *International Joint Conference on Neural Networks, 2003*, volume 3, pages 1661–1666+.

Sinka, M. P. and Corne, D. (2003a). Towards modernised and web-specific stoplists for web document analysis. In *Web Intelligence*, pages 396–404. IEEE Computer Society.

Sinka, M. P. and Corne, D. W. (2003b). Design and application of hybrid intelligent systems. chapter Evolving Better Stoplists for Document Clustering and Web Intelligence, pages 1015–1023. IOS Press, Amsterdam, The Netherlands, The Netherlands.

Wilbur, W. J. and Sirotkin, K. (1992). The automatic identification of stop words. *J. Inf. Sci.*, 18(1):45–55.

Zipf, G. K. (1935). *The Psychobiology of Language*. Houghton-Mifflin, New York, NY, USA.