# Analysis of Term Roles Along Taxonomy Nodes by Adopting Discriminant and Characteristic Capabilities

Giuliano Armano, Francesca Fanni, and Alessandro Giuliani

Department of Electrical and Electronic Engineering (DIEE)
University of Cagliari, via Marengo 2, 09123 Cagliari, Italy
{armano,alessandro.giuliani,francesca.fanni}@diee.unica.it
http://www.iascgroup.it/en/

**Abstract.** Taxonomies are becoming essential to a growing number of application, particularly for specific domains. Taxonomies, originally built by hand, have been recently focused on their automatic generation. In particular, a main issue on automatic taxonomy building regards the choice of the most suitable features. In this paper, we propose an analysis on how each feature changes its role along taxonomy nodes in a text categorization scenario, in which the features are the terms in textual documents. We deem that, in a hierarchical structure, each node should intuitively be represented with proper meaningful and discriminant terms (i.e., performing a feature selection task for each node), instead of considering a fixed feature space. To assess the discriminant power of a term, we adopt two novel metrics able to measure it. Our conjecture is that a term could significantly change its discriminant power (hence, its role) along the taxonomy levels. We perform experiments aimed at proving that a significant number of terms play different roles in each taxonomy node, giving emphasis to the usefulness of a distinct feature selection for each node. We assert that this analysis should support automatic taxonomy building approaches.

**Keywords:** Discriminant Capability, Characteristic Capability, Taxonomy

## 1 Introduction

In this paper, the underlying scenario is text categorization, where source items are textual documents (e.g., webpages, online news, scientific papers, or e-books). In particular, this work is part of a bigger project concerning the automatic taxonomy building. Taxonomies are becoming essential to a growing number of application, particularly for specific domains. They play an important role in many applications. For example, in web search, organizing domain-specific queries into hierarchies can help to better understand the queries and improve search result [13], or to improve query refinement [11]. Taxonomies, originally built by hand, have been recently focused on their automatic generation. In particular, a main issue on automatic taxonomy building regards the choice of the most suitable features (i.e., the terms in the textual documents). We deem that, in a hierarchical structure, each node should intuitively be represented with proper discriminant terms, i.e., performing a feature selection task for each node, instead of considering a fixed feature space for the entire taxonomy. To assess the discriminant power of a term, we use novel metrics able to measure it. The adopted metrics are the *discriminant capability*, that grows in accordance with the ability to distinguish a given category against others, and the *characteristic capability*, that grows in accordance to how the term is frequent and common over all categories. Our conjecture is that a term could change its role, depending on its discriminant power, along the taxonomy levels. We perform experiments aimed at analyzing such changes of role along taxonomy nodes.

The rest of the paper is organized as follows: Section 2 regards the background of this work; in Section 3 the adopted metrics are described, whereas in Section 4 the methodology of terms roles analysis is explained; experiments are reported in Section 5, and Section 6 ends the paper with the conclusions and the future work.

## 2 Background

Recently there have been a focus on the automatic taxonomy building. The motivations are obvious: manual construction is a laborious process, and the resulting taxonomy is often highly subjective, compared with taxonomies built by data-driven approaches. Furthermore, automatic approaches potentially could enable humans or even machines to understand a highly focused and fast changing domains. Several works have been devoted to taxonomy induction, in particular with respect to automatically creating a domain-specific ontology or taxonomy [7–9]. In particular, an important task is to recognize the most meaningful features. According to Luhn [6], only a relatively small number of terms, in a document, is meaningful. In fact, most terms, in a corpus, are non-informative. There are two types of terms that are meaningless for representing a topic or a category: (i) terms that occur only in a few number of documents, and (ii) terms that frequently occur in a document collection (the so-called *stopwords*); stopwords are mainly pronouns, articles, prepositions, conjunctions, some frequent verbs forms, etc. [12]. Several works in the literature focused on the analysis of stopwords in document collections [3–5, 10], proving that stopwords tend to occur in the majority of domain documents and introduce noise for IR tasks [10]. For these reasons, stopwords should be filtered in the document representation process, since they actually reduce retrieval effectiveness. Our insight is that, due to the hierarchical structure, each node intuitively should be represented with proper meaningful terms, instead of considering a fixed vocabulary for the entire structure. In other words, we deem that each document collection is unique, making useful to devise methods and algorithms able to automatically build a distinct list of meaningful features for each collection.

## 3 The Adopted Metrics

In this paper, we adopt two metrics able to provide relevant information to researchers in several IR and ML tasks. The metrics have been devised for both classifiers performance assessment and feature selection tasks [1]. As the most acknowledged approaches do not assess the discriminant power of a term, we apply the metrics for feature selection, in which they are able to evaluate the *discriminant* and *characteristic* capabilities of each feature. In particular, as the underlying scenario is text categorization, for each term the former measures the ability to distinguish a given category $C$ against others, whereas the latter measures to which extent term is pervasive in the given set of documents. The definitions of discriminant ($\delta$) and characteristic ($\varphi$) capabilities, in this scenario, are the following [1]:

$$\delta = \frac{\#(t,C)}{\#(C)} - \frac{\#(t,\bar{C})}{\#(\bar{C})} \tag{1}$$

$$\varphi = \frac{\#(t,C)}{\#(C)} - \frac{\#(\bar{t},\bar{C})}{\#(\bar{C})} \tag{2}$$

where a generic term $t$ contained in a document represents the *binary* feature under analysis, meaning that it can be assume two values, depending on the presence or absence in the document. Table 1 reports the meaning of each component in the formulas, in which the absence of term is denoted as $\bar{t}$, and the alternate class is denoted as $\bar{C}$.

Table 1: Confusion matrix entries in text classification

| | |
|---|---|
| $\#(t,C)$ | #docs of $C$ containing $t$ |
| $\#(t,\bar{C})$ | #docs of $\bar{C}$ containing $t$ |
| $\#(\bar{t},C)$ | #docs of $C$ NOT containing $t$ |
| $\#(\bar{t},\bar{C})$ | #docs of $\bar{C}$ NOT containing $t$ |
| $\#(C)$ | #docs of $C$ |
| $\#(\bar{C})$ | #docs of $\bar{C}$ |

Assuming both ranging from -1 to +1, the proposed metrics show an orthogonal behavior, and it has been proved that the $\varphi - \delta$ space is constrained by a rhomboidal shape [1], as reported in Figure 1.

## 4 Terms Roles

In this context, a term plays a distinct role in each category, depending on the rhombus region in which the term falls. Important terms for text classification

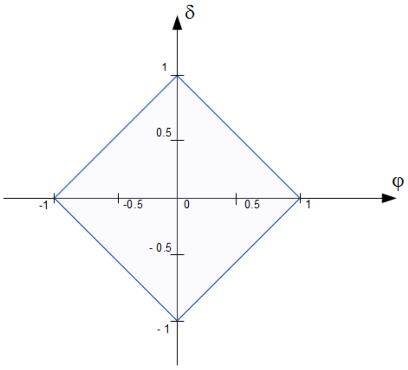

Fig. 1: The theoretical $\varphi - \delta$ space.

appear in upper and lower corner of the rhombus in Figure 1, as they have high values of $|\delta|$. In particular, a high positive value of $\delta$ means that the term frequently occurs in $C$ and is rare in $\bar{C}$; ideally, $\delta$ is $+1$ when the term occurs in all documents of $C$ and no documents of $\bar{C}$ contain it. Conversely, a high negative value of $\delta$ means that the term frequently occurs in $\bar{C}$ and is rare in $C$; ideally, $\delta = -1$ means that all documents of $\bar{C}$ contain the term, and no documents of $C$ contain it. As for the characteristic capability, terms that occur barely on the entire domain are expected to appear in the left corner of the rhombus (high negative values of $\varphi$), while stopwords are expected to appear in the right handed corner (high positive value of $\varphi$). Ideally, $\varphi = +1$ when the term occurs in each document of the entire domain, whereas $\varphi = -1$ when the term is completely absent in the domain. Figure 2 outlines the expected behavior for all cases.

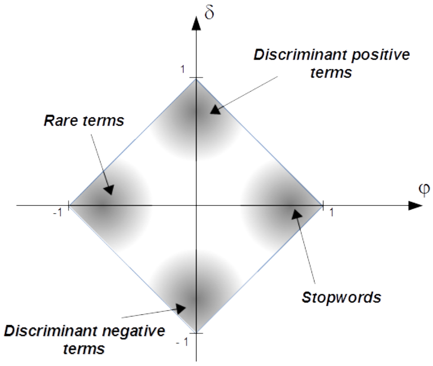

Fig. 2: The roles of terms

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 domain-dependent stopwords are located in that area [2].

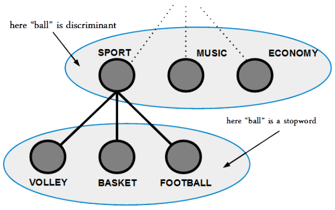

Fig. 3: Roles of the term "ball'.

The aim of this work is to analyze the roles of terms in different levels of taxonomies; in particular, we are interested in verifying how terms change their role in the ancestor nodes. For example, let us consider the domain "Sport", containing the categories "Volleyball", "Basket", and "Football"; intuitively, for the given domain, the term *ball* should be considered a domain-dependent stopword, as it is not relevant to discriminate among the cited categories. The term could change its role in the parent node. For example, if the parent node belongs to a set of siblings built with the categories "Sport", "Music", and "Economy", the term *ball* becomes, intuitively, discriminant for the category "Sport", as reported in Figure 3.

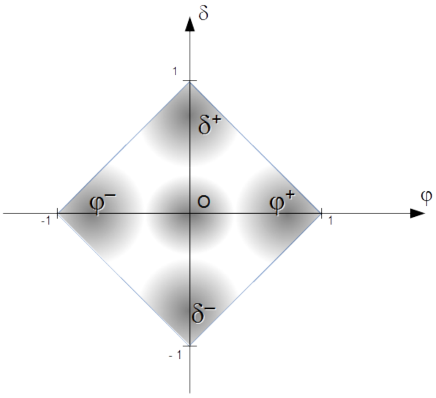

Fig. 4: The regions of the space.

In so doing, we perform a further analysis of the rhombus area defined in the $\varphi - \delta$ space. We first introduce the behavior of the metrics in the neighborhoods of origin. Theoretically, if a term has a zero value for both $\delta$ and $\varphi$ 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 $\bar{C}$ contain the term. If a term is projected close to the origin of the space, there is uncertainty in considering the term as stopword, irrelevant, or discriminant. An analysis of this region is a part of future work.

We assign the following symbols to the regions of the rhombus:

- $\delta^+$: the region in which highly positive discriminant terms are placed,
- $\delta^-$: the region in which highly negative discriminant terms are placed,
- $\varphi^+$: the region in which global and domain-dependent stopwords are placed,
- $\delta^+$: the region in which rare terms fall.
- $O$: the region of uncertainty, where we cannot actually infer the real nature of a term.

Figure 4 depicts the regions defined above. As pointed out, $O$ is the zone of uncertainty. In this preliminary study we consider $O$ constrained by a rhombus. Analytically, $O$ is defined by the equation $|\varphi|+|\delta| \leq \epsilon$; the value of $\epsilon$ identifies the dimension of the rhombus. As for the other regions, we separate them linearly. The regions defined in this work are depicted in Figure 5.

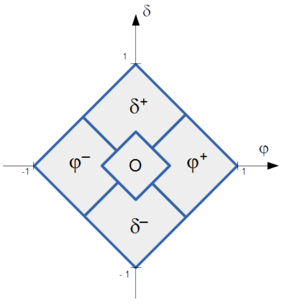

Fig. 5: The defined regions.

With the goal of better understanding the roles of terms along the taxonomy levels, we adopt a finite-state machine (FSM) representation, in which each region defined above represents a "state", whereas each transition represents the change of region: the "current state" is the region in which a term falls for a given node, and the "next state" is the region of the parent node rhombus in which the term is placed. For instance, let us consider the example of Figure 3: the term *ball* falls in the $\varphi^+$ region for the node "Basketball"; in the parent node "Sport" the term becomes discriminant and falls in the $\delta^+$ region. The associated FSM example is reported in Figure 6.

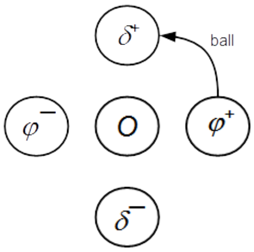

Fig. 6: Movement of the term *ball*.

## 5 Experiments

We perform experiments aimed at analyzing the term roles along taxonomy levels. In so doing, we want to verify that terms have different roles along taxonomy nodes.

### 5.1 The Adopted Dataset

Experiments are performed using a collection of webpage documents. The dataset is extracted from the DMOZ taxonomy[1] (http://www.dmoz.org). A set of 174 categories containing about 20000 documents, organized in 36 domains, has been chosen. Each domain consists in a set of siblings nodes. Aside from the leaves, each node is built with the union of the children's documents. Textual information from each page code is extracted, and each document is converted into a bag of words representation, each word being weighted with two values: $\varphi$ and $\delta$, computed by applying equations 1 and 2.

### 5.2 Analysis of Role Changes

The purpose of the following experiments is to track term movements, i.e, for each term, the change of region (a transition in the FSM model) when the focus is moved from a node to its parent. We discarded the global stopwords in this analysis, since we want to focus on domain-dependent and discriminant terms. In the following FSM charts, each edge is marked with the number of terms that participated to the associated transition. Figure 7 reports the transition for all terms in the dataset, in which the value of $\epsilon$ (i.e., the parameter that controls the size of the neutral region) is initially set to 0.1[2].

[1] DMOZ is a collection of HTML documents referenced in a Web directory developed in the Open Directory Project (ODP).

[2] We performed analyses with other values of $\epsilon$; we did not report the charts for the sake of brevity. The choice of proper shape and dimensions of the $O$ region are currently under study.

In our opinion, the transitions regarding only a few number of terms should be only due to statistical fluctuations. For the sake of clarity, we did not report, transitions marked with 1. The majority of movements, as expected, belongs to the transition $\varphi^- \rightarrow \varphi^-$, meaning that most terms (more than 99% of them) are rare or irrelevant (in accordance with the Zipf's law). This is intuitively coherent with the fact that, in a parent node, a term belongs to a more populated vocabulary; if a term is rare in a domain, it should remain rare (actually it should be more rare) in a bigger domain.

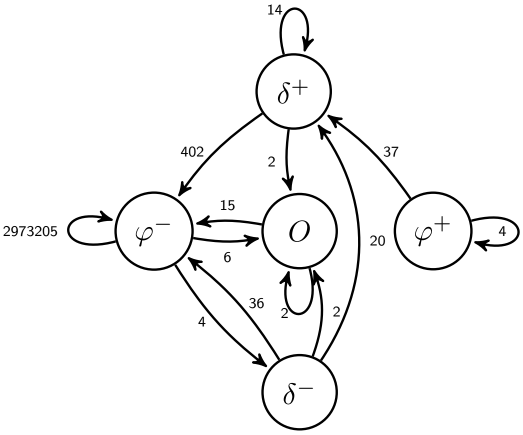

Fig. 7: FSM for the total movements with $\epsilon = 0.1$.

As for the other transitions, a term having a high value of $\varphi$ may become discriminant, as expected, in the parent node. This corresponds to the transition $\varphi^+ \rightarrow \delta^+$. Figure 7 reports 37 $\varphi^+ \rightarrow \delta^+$ transitions. This phenomenon sustains the conjecture that a domain-dependent stopword may become discriminant in the parent node. There is also a significant number of $\delta^+ \rightarrow \varphi^-$ transitions, meaning that a discriminant term (positive side) tends to be irrelevant in the parent node. As the parent node is built with the union of documents belonging to its children, there is a higher population of terms. The chosen taxonomy has a high average branching factor ($\approx 5$), hence, the frequency of a term is significantly smaller in the parent node, and the term becomes rare. The same behavior is observed in the ($\delta^- \rightarrow \varphi^-$) transition for the same reasons (the smaller number of transitions is due to a smaller number of negatively discriminant terms than the positive ones in the entire dataset).

Moreover, in the previous FSM charts, there is a significant number of $\delta^- \rightarrow \delta^+$ transitions: although, at first glance, it seems a strange behavior, it is actually not surprising: let us consider the term "ball" in the children of *Sport*, as reported in the example of Figure 8; most of them (*Volley*, *Basket*, *Football*, *Rugby*, and *Handball*) are sports played with a ball. On the other hand, there is a sibling

(*Auto Racing*) in which the term “ball” is expected to occur barely. The term is obviously negatively discriminant for *Auto Racing*, as it appears frequently in the alternate class; on the other hand, the term should be significantly frequent in the domain *Sport*, as it is expected to appear frequently in 5 siblings of 6; furthermore, looking at the siblings of *Sport*, intuitively the term should be not frequent in the alternate class, given by the union of *Music* and *Economy*. Hence, the term “ball” should be positively discriminant for *Sport*.

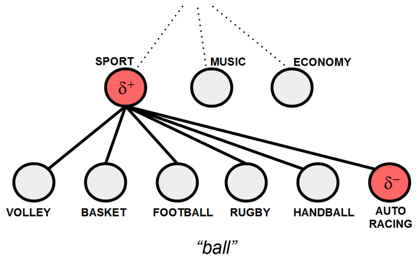

Fig. 8: Role of term *ball* in *Auto Racing* and *Sport* categories.

A further important property is that there are no $\varphi^+ \rightarrow \delta^-$ transitions; this is an expected behavior: a negatively discriminant term in the parent node ($\delta^-$ region) has very low frequency (0 occurrences in ideal case); this is in contrast with the fact that a positive characteristic term (it falls in the $\varphi^+$ region) is highly frequent in the entire domain.

Taking into account the previous analyses, we can now hypothesize that a domain-dependent stopword in a given node, probably becomes discriminant when the focus is moved in upper levels of the taxonomy; subsequently it becomes rare, and remains rare until the root of the taxonomy. Figure 9 reports the associated FSM.

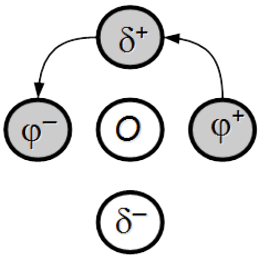

Fig. 9: The main movements of a domain-dependent stopword in a taxonomy.

As an example, let us consider the path highlighted in Figure 10; the node *Academic Department* contains several domain-dependent stopwords (that is, they belong to the $\varphi^+$ region).

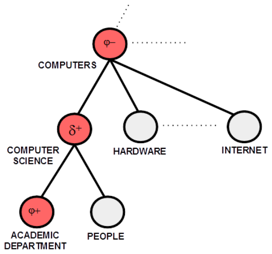

Fig. 10: Example of role change in a path.

We expect that at least some of them become discriminant in the parent node (*Computer Science*); moving up in the taxonomy path, we suppose they become irrelevant (they fall in the $\varphi^-$ region for the node *Computers*). The Figure 11 confirms this aspect, reporting the experimental results for the previous example; the node *Academic Department* contains the domain-dependent stopwords “computer”, “research”, “science”, “university”. The Figure shows the placements of these terms in the $\varphi - \delta$ space, for the given node and for the ancestors highlighted in Figure 10; it is clear how each term has the total transition $\varphi^+ \rightarrow \delta^+ \rightarrow \varphi^-$.

This behavior is an essential property of a hierarchically ordered set of documents. It is well known that in classification tasks a feature selection process improves the classifier performances; in a taxonomy, a feature (i.e., a term) may assume different roles in each node. Figure 11 clearly shows that a distinct feature selection task should be performed for each node, instead of considering a global feature space for the entire taxonomy; the methodology will be based on the selection of the most discriminant terms only, discarding irrelevant terms and stopwords (global and domain-dependent). The metrics permit to identify meaningful features to be selected in automatic taxonomy generation algorithms. Furthermore, this conjecture is the starting point of future work on hierarchical classification, in which the metrics will be adopted for performing local feature selection tasks.

## 6 Conclusion and Future Work

We proposed an analysis on how each feature changes its role along taxonomy nodes, in a text categorization scenario, in which the features are the terms in the textual documents. Results proved that a significant number of terms have different roles along taxonomy nodes, giving emphasis to the usefulness of a proper feature selection for each node. The adopted metrics permit to identify

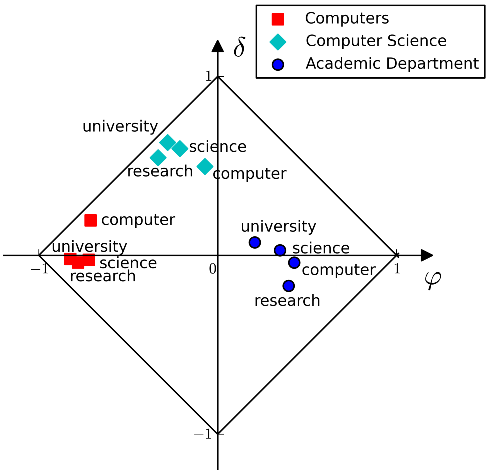

Fig. 11: Placements of terms "computer", "research", "science", and "university".

meaningful features to be selected in automatic taxonomy generation algorithms. Currently we are developing a methodology based on this metrics. Furthermore, this conjecture is the starting point of future work on hierarchical classification, in which the metrics will be adopted for performing local feature selection tasks. We are also currently planning to investigate different values of $\epsilon$ for defining the uncertainty region.

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

## References

1. Armano, G.: A direct measure of discriminant and characteristic capability for classifier building and assessment. Tech. rep., DIEE, Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy (2014), dIEE Technical Report Series
2. Armano, G., Fanni, F., Giuliani, A.: Stopwords identification by means of characteristic and discriminant analysis. In: Loiseau, S., Filipe, J., Duval, B., Van Den Herik, J. (eds.) 7th International Conference on Agents and Artificial Intelligence 2015 (ICAART 2015). pp. 353–360. SCITEPRESS Science and Technology Publications, Lisbon, Portugal (10–12 Jan 2015)
3. Francis, W.N., Kucera, H.: Frequency Analysis of English Usage: Lexicon and Grammar. Houghton Mifflin (1983)

4. Hart, G.W.: To decode short cryptograms. Commun. ACM 37(9), 102–108 (Sep 1994), http://doi.acm.org/10.1145/182987.184078
5. Kucera, H., Francis, W.N.: Computational analysis of present-day American English. Brown University Press, Providence, RI (1967)
6. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. 2(2), 159–165 (Apr 1958), http://dx.doi.org/10.1147/rd.22.0159
7. Mani, I.: Automatically inducing ontologies from corpora. In: Proceedings of CompuTerm 2004: 3rd International Workshop on Computational Terminology, COLING'2004 (2002)
8. Navigli, R., Velardi, P., Faralli, S.: A graph-based algorithm for inducing lexical taxonomies from scratch. In: Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three. pp. 1872–1877. IJCAI'11, AAAI Press (2011), http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-313
9. Poon, H., Domingos, P.: Unsupervised ontology induction from text. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 296–305. ACL '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1858681.1858712
10. Rijsbergen, C.J.V.: Information Retrieval. Butterworth-Heinemann, Newton, MA, USA, 2nd edn. (1979)
11. Sadikov, E., Madhavan, J., Wang, L., Halevy, A.: Clustering query refinements by user intent. In: Proceedings of the 19th international conference on World wide web. pp. 841–850. WWW '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1772690.1772776
12. Silva, C., Ribeiro, B.: The importance of stop word removal on recall values in text categorization. In: International Joint Conference on Neural Networks, 2003. vol. 3, pp. 1661–1666 (2003)
13. White, R.W., Bennett, P.N., Dumais, S.T.: Predicting short-term interests using activity-based search context. In: Proceedings of the 19th ACM international conference on Information and knowledge management. pp. 1009–1018. CIKM '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1871437.1871565