UNIVERSITY OF THE AEGEAN

DOCTORAL THESIS

Open-set Web Genre Identification

Author:
Dimitrios A. PRITSOS

Supervisor:
Efstathios STAMATATOS

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

at the

Dept. of Information and Communication Systems Eng.

November 13, 2019

UNIVERSITY OF THE AEGEAN

Abstract

Doctor of Philosophy

Open-set Web Genre Identification

by Dimitrios A. PRITSOS

World wide web is constantly increasing and people use information in web-pages for everyday activities. There is an emerging need for facilitating access in this huge repository in a seamless way that is in accordance with users' understanding. Genre is an important factor to characterize the properties of web-pages. Web genres (e.g., blogs, e-shop, FAQs, etc.) refer to the form, structure, and communicative purpose of web-pages rather than their topic. Web Genre Identification (WGI) provides a means to improve effectiveness of information retrieval systems by allowing sophisticated queries combining topic and genre information and ranking/grouping search results according to genre. Specialized document collections can be compiled by adopting genre-aware focused crawling. The credibility assessment of web-pages can be significantly enhanced given that information about their genre is available. Cyber-security applications like anti-phishing can also be enhanced by incorporating genre of web-pages. In case natural language technology tools should be applied to the textual part of web-pages, knowing their genre allows the selection of appropriate tools that have been trained to handle similar documents.

Existing work in WGI largely follows the closed-set classification scenario where given a genre palette and training examples for each known genre the task is to assign every new web-page to one of the known genres. However, this does not fit most of applications related to WGI. There is no consensus about the definition of a large genre palette covering most of the Web. It should be expected that large volumes of web-pages will not belong to any of the pre-defined genre labels. This could be viewed as noise in WGI. In addition, genres evolve in time, new genres emerge and existing genres are modified (e.g., blogs and micro-blogs). It seems reasonable to adopt the open-set scenario to better deal with WGI tasks. The very few existing studies focusing on open-set WGI lack an objective evaluation that will reveal their true potential.

In this thesis, we develop three open-set WGI methods. One follows the one-class classification paradigm (OCSVM) where only positive examples of a target class are used during training. Another follows the ensemble learning paradigm (RFSE) and applies random subspacing to avoid the curse of dimensionality. The third approach is a modification of k-Nearest Neighbor classifier (NNDR) that attempts to regulate the open-space risk (i.e., the area that lies away of positive examples of a class could be occupied by another, unknown, class). In addition, we examine several text representation methods including low-level and language-independent features like character n-grams and word n-grams and syntactic features like part-of-speech n-grams. We also introduce the use of distributed representations obtained by neural network language models in WGI.

Another major contribution of this thesis is the evaluation framework we propose for open-set WGI methods. In contrast to previous approaches in this field, we focus on both unstructured and structured noise. The former means that noise is composed by a random collection of web-pages without any information about their genre. The latter assumes that noise consists of web-pages of certain genres. We adopt open-set evaluation measures, variants of the well-known precision, recall, and F_1 measures, excluding true positives of the unknown class. In addition, we use graphical evaluation measures that depict the performance of the examined methods in varying conditions. We also introduce the use of the openness test in WGI studies allowing to control the homogeneity of noise and the difficulty of the task.

A series of experiments is conducted to evaluate the proposed WGI methods using the open-set evaluation framework when both unstructured and structured noise is available. The ensemble-based approach (RFSE) achieved the best overall results demonstrating its ability to handle high-dimensional and sparse representations. NNDR is significantly improved when coupled with distributed representations that provide compact and dense vectors. This method is quite competitive especially when special emphasis is put on precision rather than recall. This is important given that several WGI applications (e.g., ranking of search results) prefer to optimize precision. The one-class learning approach (OCSVM) in general is not competitive. However, it surpasses RFSE for high openness scores, that is when very few known genres are available and noise is quite heterogeneous. Several ideas for further improving the obtained results are discussed.

Contents

Abstract **iii**

1 Introduction **1**
1.1 Text Mining 1
1.2 Classifying Documents by Genre 2
1.3 Closed-set vs. Open-set Classification 4
1.4 Representation of Web-pages 6
1.5 Motivation 8
1.6 Contribution 8
1.7 Publications 10
1.8 Thesis Outline 10

2 Relevant Work **13**
2.1 Introduction 13
2.2 The Notion of Genre 14
2.3 Representation of Genre-related Information 16
2.3.1 Textual Features 16
2.3.2 Structural Features 19
2.3.3 Image-related Features 20
2.3.4 Hyperlinks and URL-based Representation 21
2.3.5 Combination of Features 23
2.3.6 Domain-specific Genre Representation 24
2.3.7 Feature Weighting and Selection 25
2.4 Machine Learning Approaches to Genre Identification 33
2.4.1 Closed-set Genre Recognition 33
2.4.2 Open-set Classification 35

Chapter 1

Introduction

1.1 Text Mining

Text mining roughly concerns knowledge discovery in texts, i.e. the process where *Information Retrieval* (IR), *Natural Language Processing* (NLP), and *Machine Learning* (ML) methods are used for extracting *high-level* information from texts. This information could refer to thematic/opinion/stylistic analysis of texts (Hotho, Nürnberger, and Paaß, 2005). Given the huge amount of texts in electronic form produced daily in Internet media, this general research field has many applications in diverse areas including business and marketing, digital humanities and cyber-security (Weiss et al., 2010).

The main tasks in text mining research are following (Aggarwal and Zhai, 2012):

- *Text Retrieval*: Given a large repository of documents, the goal is to enable easy access to the stored information by retrieving the subset of documents matching the information need of a user. A typical example is web search engines.

- *Information Extraction*: The goal is to extract specific information from documents, e.g. the names of people/places/organizations and dates of events in news stories.

- *Text Classification*: The goal is to assign labels from a predefined set to documents. Such labels could correspond to thematic area (e.g., 'politics', 'sport'), or the sentiment of texts (opinion mining) or the author of documents.

- *Text Clustering*: The goal is to group documents according to their similarity. This is used when there is no predefined list of categories and can also create structured taxonomies that organize and facilitate access to a document collection.

- *Text Visualization*: This aims at graphically depicting the main information found in a collection of documents to facilitate the exploration of similarities/differences among them and provide understandable information.

- *Document Summarization*: The goal is to provide a brief summary of a long document or a collection of documents by removing trivial details and including all crucial information. This facilitates access to collections of documents that are constantly updating.

1.2 Classifying Documents by Genre

Genre Identification is the natural progress of the almost ancient process of categorizing the human intellectual creations on such an abstract taxonomy as their Genus. Artifacts such as paintings, music pieces and written texts are always a subject of research interest to be classified based on their from, style and communicative purpose rather than their content. For example, novels or poems for documents, impressionism or expressionism for paintings, blues or funky for music, are some examples of genres that depend on structural information. Especially for documents, the defining factors for distinguishing between genres are their form, style, and communicative purpose.

There is a great debate for defining the notion of genre in the linguistic studies. Additionally, the genre notion is confusing when compared with other abstract categorizations of texts such as the *text types* or *registers* etc. Despite the methodological differences the linguistic community concluded that the idiosyncrasy of the genre taxonomy is mutable and diverse (Coutinho and Miranda, 2009). This kind of idiosyncrasy is yielded to the genre taxonomy due to the spontaneous genesis of the genre classes. Genre classes are emerging or mutating when a communication process is taking place.

Definition 1 *Genre is the genus of some arbitrary texts, which comprehensively describes their form, style and communicative purpose other than their content, where it emerges as a sociocentric interaction for accelerating the social communication when it comes to the description of the texts.*

Automated Genre Identification (AGI): Identification of the text's genre and sometime equivalent to text's register. That is the the automated identification of the form, style and communicative purpose of texts. *News* indicates a different kind of texts than *Blogs* with respect to genre. *Editorial* is different than *Article* with respect to the register while both can be considered as opinion articles written in argumentative style.

A subset of AGI is *Web Genre Identification (WGI)* focusing on the World Wide Web where enriched documents (hypertexts) are classified on a given genre taxonomy/palette (e.g., blogs, home pages, e-shops, discussion forums, etc). The ability to automatically recognize the genre of web documents can enhance performance in several applications including the following:

- IR systems can enable genre-based grouping/filtering of search results Braslavski2007,Rosso2008
  A search engine can provide its users the option to define sophisticated queries

combining genre labels and topics (e.g., blogs about machine learning or e-shops about sports equipment).

- Specialized collections and intuitive hierarchies of web page collections can be built by combining topic and genre information (De Assis et al., 2009). Genre-aware focused crawling, unlike general web-crawling, explores and downloads only relevant web-pages belonging to certain genres deAssis:2017. As a result valuable time and resources are saved and more specialized indices can be produced. The main challenge in this task is to be able to guess the genre of web-pages in advance, i.e. before the page is actually downloaded (Priyatam et al., 2013).

- Knowing the genre of web-pages can be very helpful information in order to assess their credibility in spam detection (Agrawal, Mohan, and Reddy, 2018).

- In cyber-security, genre of web-pages can be exploited to enhance anti-phishing attempts Abbasi:2015.

- The recognition of web genre can also enhance the effectiveness of processing the content of web pages in information extraction applications. For example, given that a set of web pages has to be part-of-speech tagged, appropriate models can be applied to each web page according to their genre (Nooralahzadeh, Brun, and Roux, 2014).

Despite such interesting application areas, research in WGI is relatively limited due to fundamental difficulties emerging from the genre notion itself. The most significant difficulties in the WGI domain are the following:

- There is not a consensus on the exact definition of genre (Crowston, Kwaśnik, and Rubleske, 2011).

- There is not a common genre palette that comprises all available genres and sub-genres (Santini, 2011; Mehler, Sharoff, and Santini, 2010; Mason, Shepherd, and Duffy, 2009b; Sharoff, Wu, and Markert, 2010a), moreover, genres are evolving in time since new genres are born or existing genres are modified (Boese and Howe, 2005).

- It is not clear whether a whole web page should belong to a genre or sections of the same web page can belong to different genres (Jebari, 2015; Madjarov et al., 2015).

- Style of documents is affected by both genre-related choices and author-related choices (Petrenz and Webber, 2011; Sharoff, Wu, and Markert, 2010b). As a result, it is hard to accurately distinguish between personal style characteristics and genre properties when style is quantified.

1.3 Closed-set vs. Open-set Classification

In a typical text classification task, we are given a collection of documents $\mathscr{D} = \{d_1,\ldots,d_{|\mathscr{D}|}\}$ and a set of labels $\mathscr{C} = \{c_1,\ldots c_{|\mathscr{C}|}\}$ and the task is to assign each document to some of the labels. That is, for each pair $< d_j, c_i > \in \mathscr{D} \times \mathscr{C}$ a binary answer is produced indicating whether document d_i is assigned to class c_j. Usually, text classification tasks are successfully handled by applying supervised machine learning methods (Sebastiani, 2002). This assumes the availability of a labeled training corpus $\mathscr{T} = \{d_1,\ldots,d_{|\mathscr{T}|}\} \subset \mathscr{D}$ where every pair $< d_j, c_i >$ is either a positive or a negative instance of c_i. Then, a classifier learns a function $\phi : \mathscr{T} \times \mathscr{C} \rightarrow \{True, False\}$ that approximates the target function $\check{\phi} : \mathscr{D} \times \mathscr{C} \rightarrow \{True, False\}$. The effectiveness of the classifier is estimated using another labeled dataset (test/evaluation set) $\mathscr{E} = \{d_1,\ldots,d_{|\mathscr{T}|}\} \subset \mathscr{D}$ that is non-overlapping with the training set.

Most previous studies in WGI consider the simple case where all web pages should belong to a predefined taxonomy of genres (Lim, 2005; Santini, 2007; Kanaris and Stamatatos, 2009; Jebari, 2014). This is known as closed-set classification.

Definition 2 *Closed-set Classification assumes that the training and test sets are drawn from the same distribution and all their instances necessarily belong to at least one of the predefined labels.*

There are several variations of that scenario, for example single-label (where each web-page belongs to exactly one label) or multi-label classification (where it is possible multiple labels to be assigned to a certain web-page), and soft classification (where an algorithm can return the probability score for every class from the trained label space (Geng, Huang, and Chen, 2018)).

The naive assumption of closed-set classification is not appropriate for most applications related with WGI. As already mentioned, it is not feasible to define a complete set of web genres. The scale of the Web makes any attempt to map existing web-pages to a specific genre label intractable. In addition, web genres in particular are evolving in time, some are modified or seize to exist and new ones are emerging (e.g., some years ago, blogs or tweets were unknown). The vast majority of previous work in WGI avoid to consider such concerns and as a result their effectiveness in closed-set classification conditions is over-estimated.

It is therefore realistic to assume that despite best efforts to define a long genre label list, there will always be a great amount of web-pages that do not belong to any of these. Previous work in WGI define such web-pages as *noise* (this term can also refer to the case where multiple genres co-exist and there is no dominant genre label) (Santini, 2011; Levering, Cutler, and Yu, 2008). To handle noise in WGI there are two main options. First, to adopt the closed-set classification setup having one predefined category devoted to noise. Positive training examples are given for this noise class. Since this category would comprise all web pages not belonging to the known genre labels, it would not be homogeneous and it is not clear how to sample it. Moreover, this noise class would be much more greater with respect to the other genres causing class imbalance problems.

The second option is to adopt the open-set classification setting where it is possible for some web pages not to be classified into any of the predefined genre categories (Stubbe, Ringlstetter, and Schulz, 2007; Pritsos and Stamatatos, 2013). This setup avoids the problem of class imbalance caused by numerous noisy pages and also avoids the problem of handling a diverse and highly heterogeneous class. On the other hand, open-set classification requires strong generalization with respect to the closed-set setup (Scheirer et al., 2013).

Definition 3 *Open-set Classification assumes that it is likely for samples of classes unseen during the training phase to appear in test phase. An open-set classifier should be able to accurately recognize test instances belonging to the known classes (seen during training) and avoid to be confused by instances belonging to unknown classes (not seen during training) (Geng, Huang, and Chen, 2018).*

Open-set classification is closely related to the *Novelty Detection* and *One-class Classification* where it is assumed that only positive examples of a particular class are available for the supervised learning methods. These methods have been adapted to this problem and there are several examples such as One-Class SVM, One-Class Neural Networks, etc. It might sound similar but it is not a binary classification setup for training these algorithms due to the lack of the negative examples. One-class classification requires very strong generalization and it is suitable when either the negative class is not available or it is huge and heterogeneous so that it is not possible to be adequately sampled.

It is possible to transform a (soft) closed-set classifier to an open-set one by introducing a *reject option* that is used to leave a test instance unclassified. For example, a reject option may examine how far a test instance is from the class centroids or what the difference in decision probabilities between the most likely classes is and in case some predefined criteria are not met then the test instance is left unclassified (Onan, 2018). Closed-set classification methods with a reject option are not open-set essentially since they avoid to estimate the *open-space risk*.

Each classifier attempts to draw boundaries between the known classes (i.e., seen during training phase). A closed-set classifier (no matter if it uses a reject option) separates the whole instance space by such decision boundaries. However, the samples of known classes may be gathered in specific parts of the instance space. The space far away from known class instances is known as the *open space*. The open-space risk refers to the act of labeling a test instance in the open-space (Geng, Huang, and Chen, 2018).

A more formal definition of open-set classification is one where the open space risk is considered. Let $\mathscr{T}$ be the training data, R_O the open space risk, and R_ε the empirical risk. Then the objective of open-set classification is to find a function $f \in L$ which minimizes the following *open-set risk*:

$$\arg\min_{f} \{R_O(f) + \lambda R_\varepsilon(f(\mathscr{T}))\} \tag{1.1}$$

where $f(x) > 0$ implies correct recognition and λ is a regularization constant. Thus, open-set risk balances the empirical risk and the open space risk (Geng, Huang, and Chen, 2018). In practice the empirical risk is the loss function of the open-set classification model in the training set while the open-space risk is the ratio of the open space to the full vector space.

1.4 Representation of Web-pages

In order to use supervised learning technology to WGI, it is required to transform the information in raw web documents into a quantitative representation. This means that each web-page should be represented as a numerical vector where each dimension (feature) properly captures relevant information. In addition, ideally the vectors should be dense and compact to enable ML algorithms deal with the classification task efficiently.

The web documents can be considered a super-set of the document format types because it expands Postscript [1] by introducing functionality and versatility based on HTML and virtually infinite inter-connectivity because of the hyperlinks.

In relevant literature there is a great variety of ideas aiming at document representation for WGI. The main features that can be extracted from web-pages are related to the following information:

1. The Uniform Resource Locator (URL) and hyperlinks of web-pages (and the graph formed by these connections).
2. The HTML tags and Document Object Model (DOM) structure of the web-page.
3. The textual content of the web-page.

In some cases, it has been reported that the web-pages's URL alone is sufficient for predicting its genre (Abramson and Aha, 2012; Jebari, 2014; Priyatam et al., 2013; Zhu, Zhou, and Fung, 2011). Concerning available hyperlinks in web-pages there are two parts than can provide useful information: the URL of the hyperlink itself handled as a string of characters and its *anchor text*. Alternatively, the structure of the graph which is formed by the hyperlinks and information found in neighboring pages can also be used. Usually, the neighbouring pages can contribute by amplifying the signals for the correct genre classification using either information extracted from their text or based on the assumption that pages of the same genre tend to be inter-linked (Abramson and Aha, 2012; Asheghi, Markert, and Sharoff, 2014; Jebari, 2014; Priyatam et al., 2013; Zhu, Zhou, and Fung, 2011).

[1] Postscript is the digital format used from the Desktop Publishing (e.g. PDF or PS formats). In this thesis this term is used to describe all traditional document formats such as books, magazines, newspapers, in contrast to the enriched (hyperlinked) web documents.

The HTML tags can provide useful information about the structure of web-pages. In the simplest approach, HTML tags can be treated as raw text and the frequency of specific tags is measured with some potential heuristics. However, the W3C suggested HTML web-page composition paradigm is changing and constantly violated. As a result, heuristics can only contribute but in a few practical cases. A more sophisticated and sensible approach can be the analysis of the DOM structure, where the format of the text can be captured. As an example, e-shop web-pages are different from the academic web-pages. This resembles the difference in typographic format of a printed magazine and a printed newspaper. However, most likely several heuristics are needed for identifying these structures, because of the HTML composition paradigm violation (Mehler and Waltinger, 2011).

The bulk of research work in WGI has focused mostly on the features which can be extracted from the textual part of web-pages (i.e., after the removal of HTML tags) (Mason, Shepherd, and Duffy, 2009c; Sharoff, Wu, and Markert, 2010a; Sharoff, Wu, and Markert, 2010b; Nooralahzadeh, Brun, and Roux, 2014; Onan, 2018). The following are the main categories of textual features:

1. Lexical features: Each web-page is seen as a series of tokens and frequencies of specific words (e.g. function words) or sequences of tokens (e.g., word n-grams) can be measured. In addition, information about the length of words and sentences can be useful.

2. Character features: Each web-page is handled as a alphanumeric string and usually frequencies of character n-grams can provide a very detailed and highly dimensional representation.

3. Syntactic features: This requires some kind of sophisticated analysis by NLP tools that can provide information about the syntactic patterns found in the web-pages. One popular and relatively simple approach is the use of part-of-speech (POS) n-grams. Syntactic features are language-dependent and their reliability correlates with the error rate of the used NLP tools.

Typical term weighting schemes, like binary, Term Frequency (TF) and Term Frequency - Inverted Document Frequency (TF-IDF) are popular in WGI. In addition, there are some schemes specifically designed for WGI tasks like *Term Frequency - Inverted Genre Frequency* (TF-IGF). This is an extension of TF-IDF that is based on the frequencies of a term in the documents of particular genre rather than the whole corpus (Sugiyanto et al., 2014).

Recently, *distributed representations* provide an alternative way to represent documents using neural network language models mikolov2013distributed,le2014distributed. In contrast to the popular n-gram features that produce sparse vectors, distributed representations produce dense vectors of relatively low dimensionality. This approach has obtained state-of-the-art effectiveness in several text classification tasks but it has not thoroughly tested in WGI so far.

1.5 Motivation

As already mentioned, the vast majority of previous work in WGI adopt the closed-set classification scenario that is not realistic and leads to an over-estimation of performance. Since it is not feasible to define a complete list of genre labels and genres constantly evolve in time, the open-set classification scenario better suits WGI.

Among the few attempts to follow open-set classification in WGI, very few use pure open-set classifiers stubbe2007genre,Asheghi2015. An additional issue is how to handle the test web-pages belonging to unknown genres. One option is to consider these as *unstructured noise* where the true genre of noisy pages is not available and another is to examine *structured noise* where the true genre of noisy pages is available (yet unknown during the training phase).

So far, it is not clear what specific open-set classification methods can better handle these cases. In addition, there is lack of a evaluation framework that can appropriately measure the effectiveness of open-set WGI methods with the presence of either unstructured or structured noise. This requires the use of appropriately defined evaluation measures and the suitable design of experimental setup. In addition, we need a clear way to compare different methods in application-dependent conditions where, for example, precision may be considered more important than recall.

Most previous studies attempt to combine heterogeneous information coming from the hyperlinks between web-pages, the HTML code and the textual content of web-pages. Despite the usefulness of all these information, the main question is whether it is possible to accurately predict the genre of a web-page focusing on its textual content since this is not affected by technology changes and habits of web developers or arbitrary changes in neighboring web-pages.

There is a great variety of text representation measures applied to WGI, most of them attempt to capture the stylistic properties of web genres. It is not yet clear how specific approaches, like word and character n-grams, known to be very effective in closed-set WGI (Sharoff, Wu, and Markert, 2010a), are still effective in open-set WGI where the dimensionality of the representation may severely affect the ability of the open-set classifier for generalization.

Finally, the recent success of the use of distributed representations acquired by neural network language models in other text classification tasks is a strong motivation to attempt to examine their effectiveness also in open-set WGI. One main advantage of such approaches is that they produce a space of relatively low dimensionality and in theory this may be an advantage for specific open-set classifiers that may suffer when irrelevant and redundant features are available.

1.6 Contribution

This thesis focuses on open-set WGI and examines specific algorithms and experimental setups that allow their evaluation in realistic conditions. More specifically, the main contributions are listed bellow:

- An approach based on one-class classification, where only positive training examples of a target class are considered, is introduced to WGI. The proposed method is based on *one-class support vector machines* (OCSVM) and is modified to handle multi-class open-set classification. This algorithm is presented in detail in section **??**.

- The *Random Feature Subspacing Ensemble* (RFSE) is introduced to WGI. This open-set classifier is based on an existing approach originally proposed for authorship attribution and it is adopted to better handle the WGI task (Koppel, Schler, and Argamon, 2011). This algorithm has been implemented in python and in its general form can handle any kind of text representation[2]. This algorithm is presented in detail in section **??**.

- Another open-set classifier, the *Nearest Neighbors Distance Ratio* (NNDR) is introduced to WGI. This is a modification of the well-known k-Nearest Neighbor classifier (Mendes Júnior et al., 2016) and it is extended to better suit the WGI requirements. This algorithm has been implemented in python[3] and is presented in detail in section **??**.

- The noise (i.e., web-pages not belonging to any of the known genres) in WGI is distinguished into *unstructured* and *structured* noise and each case is thoroughly studied. The former considers all unknown genres as a common heterogeneous class. The latter admits that there is structure in the unknown web-pages, namely the existence of genre labels not seen during the training phase. In this thesis it is introduced the *openness* as an indication of how the number of known classes is compared to the number of unknown classes. This concept is borrowed by relevant work in visual object recognition (Scheirer et al., 2013) and it perfectly suits the WGI task.

- An experimental framework suitable for evaluating open-set WGI algorithms is introduced including abilities to study different kinds of noise (unstructured or structured). The use of openess enables the study of open-set WGI where the difficulty of the task is explicitly controlled (i.e., few known classes vs. many unknown classes or many known classes vs. few unknown classes). In addition, appropriate evaluation measures provide a detailed view on the obtained performance. This is especially important since evaluation measures usually involved in closed-set classification can be misleading since they handle all classes equally. However, in open-set WGI, the class of unknown web-pages (including all web-pages that do not belong to known genres) is usually much larger than the known classes and it should be treated in a special way as it is explained in Chapter **??**.

[2] https://github.com/dpritsos/RFSE

[3] https://github.com/dpritsos/OpenNNDR

- The proposed open-set WGI algorithms are extensively evaluated using the aforementioned experimentation framework. The particular hyper-parameters and settings that allow these algorithms to achieve as good results as possible are examined. In addition, the use of different kinds of text representation is considered and their effect on the performance of each algorithm is studied. The most popular textual features in WGI covering lexical, character, and syntactic features are considered.

- The application of distributed representations acquired from neural network language models in WGI is explored. The effect of such low dimensional and dense representations on the effectiveness of the NNDR open-set WGI algorithms is studied. It is demonstrated that especially the precision of this approach can be considerably enhanced making it more suitable for specific WGI applications.

1.7 Publications

Parts of the work described in this thesis have already been published in scientific journals and conference proceedings. The list of related publications is following:

- D.A. Pritsos, and E. Stamatatos, Open-set Classification for Automated Genre Identification, In *Proc. of the European Conference on Information Retrieval* (ECIR 2019), pp. 207-217, LNCS 7814, Springer, 2013.

- D. Pritsos and E. Stamatatos, The Impact of Noise in Web Genre Identification, In *Proc. of the International Conference of the Cross-Language Evaluation Forum for European Languages* (CLEF 2015), pp. 268-273, LNCS 9283, Springer, 2015.

- D. Pritsos and E. Stamatatos, Open Set Evaluation of Web Genre Identification, *Language Resources and Evaluation*, 52(4), pp. 949-968, Springer, 2018.

- D. Pritsos, A. Rocha, and E. Stamatatos, Open-Set Web Genre Identification Using Distributional Features and Nearest Neighbors Distance Ratio, In *Proc. of the European Conference on Information Retrieval* (ECIR 2013), pp. 3-11, LNCS 11438, Springer, 2019.

1.8 Thesis Outline

The rest of this thesis is outlined below.

Chapter 2 discusses relevant work on AGI and WGI tasks. Definitions and uses of genre from the fields of linguistics and computational linguistics are presented. The state-of-the art for the representation of web-pages and the ML methodologies for genre identification are discussed. The few open-set WGI approaches are described.

Finally, the available corpora for evaluating WGI methods and their properties are discussed.

Chapter **??** focuses on open-set WGI and analytically presents the three algorithms examined in this thesis (i.e., OCSVM, RFSE, and NNDR). The characteristics of these methods and their differences of with existing approaches are discussed.

Chapter **??** introduces the experimental framework proposed in this thesis for evaluating open-set WGI approaches. The use of openess as a means to control the difficulty of WGI tasks is discussed. Appropriate evaluation measures are defined for both unstructured and structured noise.

Chapter **??** deals with the experimental analysis of OCSVM and RFSE algorithms. The evaluation corpora used in this study and their properties are discussed. Experiments when structured and unstructured noise is considered are presented. The effect of text representation on the effectiveness of the examined methods is studied.

In Chapter **??**, the usefulness of distributed representation in open-set WGI is presented. The NNDR algorithm is evaluated using traditional n-gram-based features and distributed features. Experimental results show how the performance of this algorithm is affected and it compares with OCSVM and RFSE.

Finally, Chapter **??** summarizes the main conclusions drawn from this study and discusses future work directions.

Chapter 2

Relevant Work

2.1 Introduction

This chapter describes previous work in genre recognition. First, the notion of genre is discussed using approaches from different disciplines and background. Important aspects of genre are noted and a general definition that is adopted in this study is provided.

In general, genre recognition is viewed as a text classification task. Thus, the main issues that are studied are the following:

- Represent documents in a feature space.
- Learn a model that can distinguish between classes.

Genre-related information can be extracted from various sources. Since genre is mainly associated with form, structure, and communicative purpose of documents, features can relate to textual content, visual appearance, URL and graph of interlined web-pages, etc. In addition, as concerns textual features, information about style is far more important than topic of documents. The existing approaches to define suitable representations are analytically described. We include in this discussion both AGI and WGI tasks.

There is also a great variety of classification algorithms applied to genre recognition tasks. These include general-purpose ML methods and approaches specifically-built for these tasks. Special emphasis is given in the type of classification setup adopted by existing approaches, mainly whether a closed-set or an open-set scenario is followed.

Finally, we present an overview of existing resources to evaluate WGI approaches. A list of corpora used in previous studies and their main characteristics are described.

2.2 The Notion of Genre

In general, genre is related to form and communicative purpose of texts rather than their theme. It is closely related to style and *Genus*[1] (Sugiyanto et al., 2014). Approaches to define text genre start mainly from two directions: linguistics and computational analysis of language (e.g. computational linguistics, natural language processing, text mining).

In studies of linguistics there is a great debate in defining the notion of genre as an abstract categorization scheme of texts and the relations between them. Despite the methodological differences the linguistic community concluded that the idiosyncrasy of the *genre taxonomy* is mutable and diverse (Coutinho and Miranda, 2009). This kind of idiosyncrasy is yielded to the genre taxonomy due to the spontaneous genesis of the genre classes. The genesis of a genre class is a socio-centric interaction which is emerging from the need to describe the texts in order to accelerate the social communication procedure. Thus, genre classes are spontaneously emerging while the communication procedure is taking place.

Humans can efficiently recognize the genre-types by processing the texts intuitively. However, there is a lack of consensus for defining genres, particularly when specific names (labels) should be assigned to the genres. There there was an effort of several user studies for eliciting the mechanics in the process of genre identification and tagging. The results on user agreement were very discouraging. Also, when humans attempt to describe specifically the terms or/and the attributes which they use to identify different genres, there is a great confusion and disagreement. A convincing explanation for this is the plethora of textual, stylistic and conceptual description terms which humans use and depend on their background (e.g., teachers, scientists or engineers use different vocabularies to describe texts belonging to a common genre (Roussinov et al., 2001; Crowston, Kwaśnik, and Rubleske, 2011).

Researchers from cognitive science found that humans are recognizing the genre type of a document (or web-page) using cognitive processes related mostly to the form of the text. Particularly they used configured apparatus for tracking the eyes movement while subjects attempt to recognize genre of documents. One can resemble the process like navigation where the eyes are constantly moving while they are focusing for small fragments of time in landmarks of interest. The pausing of the eyes on the text "landmarks" is called *fixation* while the "jumping" movements of the eyes is called *saccadic*. The whole process aimed to locate information of interest such as specific text forms, names, verbs, or phrases that are related to the abstract concept in order to decide whether the text matches their interest and is worth of further reading. They systematically found that the process of finding the genre-type of the text is the same as to find out whether a text id worth of further reading. Thus, the knowledge of a genre taxonomy definitely accelerates the communication procedure and helps readers of the text to find the information of interest faster (Clark et al., 2014).

[1] Genus in Greek means *type* or *class*

The discipline of the *English for Academic Purposes* (EAP) has vividly discussed the divergence in the genre taxonomies between the different academic disciplines and reasoned the utility of the genre taxonomy for enabling the teachers and the students to improve their rhetorical and written language skills with the purpose of improving the teaching procedure. What is important to note for this study is the conclusion that any given certain genre conveys information about the communication purpose of the document, i.e. as text identity carrier, but it can also contain the same style and other language properties when the purpose is similar. For example, the article of newspaper and an article from a magazine can be claimed to belong to different genres although they are mainly governed by the same linguistic properties. Therefore, for the witter of a text is is very important to be aware (thus to be taught) of the different genres and the taxonomy of genres in order the text (s)he produces to be recognizable by the reader (Hardy and Friginal, 2016; Melissourgou and Frantzi, 2017; Al-Khasawneh, 2017). However, genre itself requires different level of human reading abilities to be recognized and even with these skills different humans may disagree (McCarthy et al., 2009).

The utility of text genre identification has been realized by the journalism professionals. There are well-defined structures and guidelines given by newspaper editors about how to present, e.g. news articles. The structure consists of abstract elements and they follow specific paradigms, like the *inverted pyramid* (i.e., contents are structured from the most important to the least important information), *Martini Glass* (i.e., it first presents a summary of the story, then an inverted pyramid and finally a chronological elaboratio), *Kabob* (i.e., it starts with an anecdote, continues with the main story and closes with a general discussion) and *Narrative* (i.e., it presents a chronological sequence of events) (Dai, Taneja, and Huang, 2018).

Some terms used in relevant literature, like *register*, and *text type* seem very relent to genre. Actually, they are used interchangeably, complimentary and even contradictory (Melissourgou and Frantzi, 2017). Although the exact definitions of these terms deviate according to the scholar and their background, text type is generally associated with linguistic properties of documents. Register usually refers to non-linguistic terms like the purpose of communication, the relation between speaker and hearer etc. Genre can be viewed as more general than both text type and register since it combines linguistic and non-linguistic information.

From a computational analysis point of view, genre (and genre taxonomy) is important as a classification factor to distinguish between documents. Genre labels are defined according to their association with practical applications rather than based on a rigid theoretical background (Kanaris and Stamatatos, 2009; Santini, 2007). Genre identification is a style-based text categorization task. Another similar task is authorship attribution where the focus is on identifying the *personal style* of the author (Stamatatos, 2009; Koppel, Schler, and Argamon, 2011; Koppel and Winter, 2014). On the other hand, genre is mainly regarded as a *group style*. Foe example scientists use a common form of language to write research papers, journalists describe news events and their opinion using similar patterns, bloggers express their beliefs and interests based on similar structures, etc.

As concerns web genres (and their respective taxonomy), the utilities and opportunities that can provide as well as the difficulties they impose have been eloquently analyzed. It has been pointed out that the genre taxonomy summarizes the type and style of texts in a single term as a communicative act (De Assis et al., 2009). In the domain of WGI, usually a web genre palette is defined usually obtained from a top-down approach, where a group of domain-experts design the taxonomy based on specific objectives of the task (Crowston, Kwaśnik, and Rubleske, 2011). Moreover, the genre palette may flat or hierarchically-structured (Wu, Markert, and Sharoff, 2010). The former assumes that genre labels are independent while the latter defines a hierarchy of genres and sub-genres. Another important issue is whether a web-page should belong to exactly one genre label or page segmentation should be applied first and then each segment should be assigned to a genre label (Madjarov et al., 2015; Jebari, 2015).

As described so far, there is agreement for the criteria which are defining the genres (and web genres) in a given domain. These are, the style, form, and the communicative purpose of documents. In theory, topic is considered orthogonal to genre. However, thematic information can also be useful in automated genre identification. For example, the genre of academic home web-pages is distinguished by a specific vocabulary. The genre of research papers also use specific science-related terms. Certainly, some of these terms may be too specific (e.g. about biology, mathematics, or computer science). However, content-specific information can be used to differentiate scientific documents from non-scientific documents (Coutinho and Miranda, 2009; Crowston, Kwaśnik, and Rubleske, 2011; Kanaris and Stamatatos, 2009; Jebari, 2015; Gollapalli et al., 2011).

Considering the above discussion, it is clear that the notion of web genre depends on the use of this information. In this thesis, our approach is influenced by the use of web genres as a classification factor in order to enhance the potential of information retrieval systems. In particular, we use the following definition:

Definition 4 *A web genre is a class of web documents that share form, structure, and communicative purpose properties. Every web-page is always derived under a unique class distribution and the class distributions are not overlapped.*

2.3 Representation of Genre-related Information

2.3.1 Textual Features

The textual content of a document is the most analyzed source of text-related information. Similarly, the textual part of a web-page is considered very important in WGI studies (Mason, Shepherd, and Duffy, 2009a; Sharoff, Wu, and Markert, 2010b). As it has already been explained, style rather than topic is crucial in genre recognition. However, it is not clear how style properties of documents can be captured adequately. In addition, style is affected by both genre and the personal style of the author. Ideally, the extracted measures should only depend on the former.

There is a great variety of textual features than can be extracted from documents and be used in genre recognition (Kanaris and Stamatatos, 2009; Kumari, Reddy, and Fatima, 2014; Levering, Cutler, and Yu, 2008; Lim, 2005; Mason, Shepherd, and Duffy, 2009b; Onan, 2018; Petrenz and Webber, 2011; Sharoff, Wu, and Markert, 2010a; Nooralahzadeh, Brun, and Roux, 2014). The main categories of such features are described below.

One simple way to represent documents is based on n-grams of either words or characters. This is a language-independent approach and has been demonstrated to be quite effective in WGI studies (Kanaris and Stamatatos, 2009; Sharoff, Wu, and Markert, 2010a; Kumari, Reddy, and Fatima, 2014). In addition, surface features that are considered important to quantify stylistic properties of documents, such as statistics (i.e., count, mean, max, etc.) of word length (in characters), sentence length (in words), paragraph length (in words), capitalized word, lowercase word, punctuation marks, type/token ratio etc. (Feldman et al., 2009; Santini, 2005; Onan, 2018). All these features attempt to represent information operating on lexical or character level.

Another popular idea is to attempt to quantify the difficulty of understanding the information included in documents by using *readability assessment* features. The main purpose of developing such features is to help in the evaluation of a text with respect to measure the degree of comprehension by the reader. Examples of readability assessment features are the word variation index (OVIX), the nominal ratio (NR) and LIX (Falkenjack, Mühlenbock, and Jönsson, 2013):

$$LIX = \frac{A}{B} + \frac{C \cdot 100}{A} \tag{2.1}$$

where A is the number of words, B is the number of special characters (i.e., colon, period, capital fist letter), and C is the number of long words (more than 6 letters for the English language).

A more sophisticated type of features concerns the syntactic properties of documents since the grammar of sentences is considered important for stylistic purposes (Sharoff, Wu, and Markert, 2010a; Petrenz and Webber, 2011). Moreover, this information is less likely to depend on topic of documents in comparison to lexical and character features. The simplest form of capturing syntactic information is the use of part-of-speech (POS) n-grams where the texts are analyzed by a POS tagger that assigns a tag in each word and then sequences of POS tags are counted. Other syntactic features are based on a more elaborate analysis of documents by NLP tools, like full syntactic parsers. Examples of such syntactic features include average dependency distance, ratio of dependencies, sentence depth (in dependency terms), unigram dependency type (based on token terms), average verbal arity, unigram verbal arity, tokens per clause, number of prepositional components, etc (Falkenjack, Mühlenbock, and Jönsson, 2013; Falkenjack, Santini, and Jönsson, 2016). A major weakness of such features is that their usefulness depend on the accuracy of the NLP tools used to extract them from documents (Stamatatos, 2009). This is especially crucial in case the documents that have been used for training the NLP tools significantly differ from the documents we want to analyze.

A text is usually viewed as a sequence of words or characters. However, an alternative idea is to construct a graph from a document and then use graph metrics to represent the properties of documents. Such graph-based features are discussed in (Nabhan and Shaalan, 2016) aiming to enhance effectiveness in genre recognition. An unweighted graph is built from each document based on word bigrams found within sentence boundaries. Each word is a node of the graph and if a bigram is found in the text an edge connects the respective words. The frequency of bigram was not taken into account.

Then, graph-based measures are extracted to represent documents including node degree, clustering coefficient, average shortest path length, network diameter, number of connected components, average neighborhood connectivity, network centralization and network heterogeneity. The average node degree, i.e. the number of neighbor connections, shown to be an important criterion for discriminating for example scientific to humorous web-pages. A higher average of node degree may indicate a preference to use an established vocabulary.

A high value of clustering coefficient would mean there is tendency for a set of nodes to cohere or stay connected in a sub-network. The Religion, Fiction, and Adventure classes seem to have relatively high value of clustering coefficient as compared to News, Editorial and Hobbies. A high number of connected components indicates topic diversity within a genre. News and Hobbies have shown to have higher score, i.e. higher diversity, than Religion and Fiction. In addition, a relatively high score in network Centralization seems to be a good indicator for Fiction and Adventure genres. The network heterogeneity was found to be higher in News and Hobbies and this reflects the tendency of the graph to have links between high-degree to low degree-nodes. This can indicate a tendency to use function words in text. Genre-specific graph characteristics also found in that study (Nabhan and Shaalan, 2016) including high global clustering coefficient found for Learned and Religious text genres. Moreover, average local clustering strongly correlates to the node degree shown to be a good indicator for genres showing concentration to specific concepts.

Finally, the graph-based measures can also be used for discovering the existence of sub-genre within a genre such as in News. It has been shown that there are some areas within the News genre where the bigram graph has high node connection concentration (or high edge concentration).

In (Kim and Ross, 2010) the *Harmonic Descriptor Representation* (HDR) of web-pages is proposed. This is inspired by the musical analogy of a string of a musical instrument. Each document is consider to be a temporal sequence of symbols (i.e. characters or words). Particularly, instead of counting the overall frequency of terms, the intervals of the the occurrences of terms within the document are measured. This shows how the occurrences of a term are distributed within a document.

This approach defines *Range* as the interval between the initial an the ultimate occurrence of the term in a document and *Period* as the time duration (i.e. the count of characters) between two consecutive occurrences of the term. Then HDR word encoding is a tuple of three explicit measurements defined as follows:

1. FP is the time duration before the first occurrence of the symbol in a document (i.e., the period before the first occurrence divided by the total number of characters into the document).

2. LP is the time duration after the last occurrence of the symbol (i.e., the period after the last occurrence divided by the total number of characters)

3. AP is the average period ratio calculated as follows:

$$AP(s,d) = \begin{cases} \frac{|d|-(f_s+1)}{max(P_s)(|d|-(f_s+1))}, max(P_s) > 0 \\ 1, max(P_s) = 0 \end{cases} \tag{2.2}$$

   where f_s is the frequency of symbol s in document d, P_s is the set of periods between all consecutive occurrences of s in d and $|d|$ is the length in characters of d.

2.3.2 Structural Features

As already discussed, genre is mainly associated with form of the presented information. However, it is quite unclear how this information can be quantified appropriately. The easiest way is to focus on HTML tags by counting the HTML tags frequency in the hypertext (Kanaris and Stamatatos, 2009). Special focus in some cases is given to the image tags and the hyperlink tags (Lim, 2005; Levering, Cutler, and Yu, 2008). These sources of information are useful and usually their combination with textual features enhances the performance of WGI model. In addition there are very few cases where the DOM object structure is analyzed for extracting information but usually as part of the whole set of features selected and not as a stand alone choice (Mehler and Waltinger, 2011). Another interesting approach is to view a web-page as an image and attempt to extract visual features that describe what components are found and in what position (Levering, Cutler, and Yu, 2008).

There are also other cases where only pure structural information of a web page, i.e. the HTML tags, are exploited [Philipp Scholl].

Structure indicative features have also been combined with SVM for the WGI task, specifically for the case of *News article* sub-genre identification. Experimental results show that reasonable performance, although, this kind of features are importing even more issues. At first are difficulty to be captured for example counting the HTML tags or by analyzing the HTML DOM tree from a browser is the best practice to follow. Moreover, this kind of information usually is vague and small (Cortes and Vapnik, 1995) .

An approach that is based on structural features is presented in (Mehler and Waltinger, 2011). They focus on the web genre of homepages and its sub-genres (i.e., personal, conference, project). The web-pages are first automatically segmented into their constituent parts (e.g., for the personal academic homepage the segments are: contact information, personal information, publications, research, and teaching). Then, each page is represented according to the detected segments that were found in it. The reported results show a significant increase in performance when this structure-based method is compared with traditional approaches based only on textual features.

2.3.3 Image-related Features

In (Chen et al., 2012) there is a very interesting approach where image processing features have been used in a AGI task applied to office documents. In their experiments, interestingly they also used image-based features that were found significantly better that regular textual features when comparing their work to previous ones. The combination of both kinds of features increased the performance even more.

The image-based features were extracted by splitting the image of the document into 25 tils (5 horizontally and 5 vertically) plus a full-page til. The features used were: (a) *Image Density*, (b) *Horizontal projection*, (c) *Vertical projection*, (d) *Color correlogram*, (e) *Lines*, (f) *Image size*. In all cases the document images where converted to black and white for these features to be extracted. The exception is the correlogram which analyzed the full color spectrum of the document in its image format. The image-based features described above are similar to the ones used in (Clark et al., 2014).

- The mage density utility was used for differentiating where the images and the text were located. In addition the titles from the rest of the text could be also separated. To capture this feature the black to total pixels ratio was calculated for each til of the document.

- The horizontal projection was used for differentiating the slides where the text is large and less than the rest of the non-slides documents. After the process required for locating the text boxes (similarly tho the OCR software) then a five-bin histogram were used for identifying the majority of the text font sizes.

- The vertical projection was used to differentiate the papers from tables by capturing the number of text columns and the distribution of their width. Similarly to the horizontal projection a five-bin histogram of column width were used.

- The color correlogram represents the spatial correlation of colors. The process is starting by quantizing the colors to a 96 scale in distance range for 0 to 1. In addition 3 pixels are used thus every til of the document has 288 dimensions.

The selection of the optimal features for reducing even further the dimensions was operated using the *Maximally Relevant Minimally Redundant* (mRMR) method, resulting 50 features per til. The preservation of the location of the spatial color correlation coefficients is important thus an implicit strategy was followed. Particularly after the mRMR the selected features where preserved to their til-vector position and then all tils vectors concatenated into one vector. Finally the non-selected features from mRMR where discarded and the "compressed" form of the concatenated vector was the final outcome of the correlogram preprocessing.

- The lines were used particularly for locating tables. The process was operated on the full-page til and it was measuring the continuous sequence of black pixels of the black and white form of the picture. Then a line-length histogram was used for discriminating the table lines from other lines present in a text such as header of footer lines often met in textbooks.
- The image size was operated only on the full-page size, for finding the page size of the document and differentiate the papers form slides or picture usually having different sized while papers usually delivered in a specific size page size.

Their reported experiments of that study were conducted to a very special case of the AGI research and for a very specialized taxonomy of office documents. The corpus included papers in PDF format, photos in JPG format, PowerPoint slides, and tables in documents. This corpus has been collected manually and then also manually annotated. *Fleiss' Kappa* agreement score for the annotators, has been used in order to evaluate the quality of their corpus (the *Kappa* score was from 0.88 to 0.92).

2.3.4 Hyperlinks and URL-based Representation

The web is structured as a directed graph where each web-page is linked with other pages through hyperlinks. Information about incoming and outgoing hyperlinks is important for WGI. In addition, information found in web-pages that are linked with the one in question could also be used.

In addition, each web-page has a unique address, the *Uniform Resource Locator* (URL) that is used to identify it. Usually, important information is encoded in URLs and sometimes this may refer to genre. For example, the string "blog" is quite likely to appear in a the URL of blogs. Several previous studies attempt to exploit this kind of information.

To begin with, a study is based on the web-graph and the implicit genre relation among web pages assuming that neighbouring web pages are more likely to belong to the same genre, a property called *homophily*. Then, the content of neighboring pages are used to enhance the representation of a given web page in a semi-supervised learning framework (Asheghi, Markert, and Sharoff, 2014) (More details to be written here).

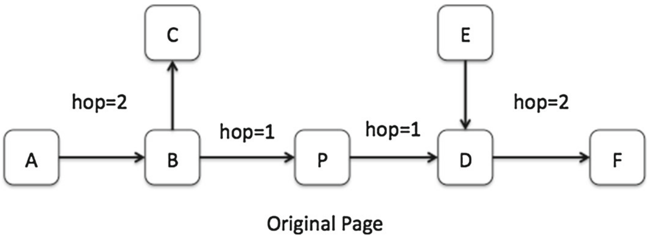

FIGURE 2.1: A directed graph of web-pages (Zhu et al., 2016).

GenreSim is a link-based graph model which exploits link structure to select relevant neighbouring pages in order to amplify the information required for a page to be classified to a genre taxonomy. This algorithm improves performance of WGI significantly in cases where the textual information is very limited in a web-page such as movie homepages, photography websites etc. On the other hand, the reported experimental results indicate that in regular web-pages, where the textual consists of at least a couple of paragraphs, the advantage of using hyperlink-based graph information is not remarkable (Zhu, Zhou, and Fung, 2011; Zhu et al., 2016).

GenreSim is a ranking algorithm based on *PageSim* algorithm, extended to fit in the problem of WGI. Similar to all this kind of algorithms, is based on the assumption that the more web-pages referred to a particular page, the more this page is related to them with respect to topic and/or genre. As concerns genre class, GenreSim focuses on *forward* $F(p)$ and *backwards* $B(p)$ hyperlinks. Moreover, utilizing the entire graph structure, web-pages are characterized as *Hubs* $H(p)$ or *Authorities* $A(p)$. The null hypothesis of the algorithm is that the web pages of the same genre are inter-connected with their hyperlinks. Consequently, a few pages backwards and forwards to a specific web-page compose a small network of the same genre. Using this "genre-network", the textual (and partially the structural) information of neighbouring web-pages can be used to amplify the signals required to classify a new web-page to that genre.

In more detail, hubs are pages with many outgoing hyperlinks, whereas pages with many incoming hyperlinks are called authorities. The number of incoming and outgoing hyperlinks are increasing the respective scores as shown in equation 2.3. However, web-pages with high score but with few backward hyperlinks are quite likely to be *spam* pages. In order to regulate this, the $\omega(p)$ factor is introduced in equation 2.4, to reduce the score for the web pages with few backward hyperlinks. In addition, this is also useful to normalize the few links issue. That is, the number of the backward links is correlated to the number of links the page itself contains.

$$\begin{aligned} H(p) &= \textstyle\sum_{u \in V|p \to u} \omega(p) A(u) \\ A(p) &= \textstyle\sum_{v \in V|v \to v} \omega(p) H(u) \end{aligned} \tag{2.3}$$

$$\omega(p) = \frac{N}{|\log N - \log N(p)| + 1} \tag{2.4}$$

Therefore, the score for a new web-page in a given G graph of web-pages, is calculated by equation 2.5. In general, the genre-selection recommendation score is propagated to the graph path $P(u,v)$ as indicated by the $Score(u,v)$ function of equation 2.6. Therefore, the score of a recommended web-page is decreasing gradually as this pages lies away (in hops) from the web-page to be classified. The d factor is set to be 0.5, i.e. the page score is decreasing by half for every hop away from the page under examination (see Figure **??**).

$$Score(p) = H(p) + A(p) \tag{2.5}$$

$$Score(u,v) = \begin{cases} \sum_{p \in P(u,v)} \frac{dScore(u)}{\prod_{x \in p, x \neq v}(|F(x)|+|B(x)|)}, & v \neq u \\ Score(u), & v = u \end{cases} \tag{2.6}$$

Finally, the similarity of the candidate neighbour pages to the one under evaluation is based on the ratio of the min and the max path-score sums of all the possible paths, backwards and forwards, to the page under evaluation. This is defined as follows:

$$Sim(u,v) = \frac{\sum_{i=1}^{n} min(Score(v_i,u), Score(v_i,v))}{\sum_{i=1}^{n} man(Score(v_i,u), Score(v_i,v))} \tag{2.7}$$

Hyperlinks themselves can be exploited by extracting information from the URL string and not from the hyperlink-graph. Particularly, a URL can be segmented to its components, i.e. the domain name, the path after the domain and the anchor text. Special characters such as $\{,.,?,\$,\%\}$, top-level domains $\{.gr,.uk,.com,etc\}$, and file suffixes such as ".html", ".pdf" are usually discarded and then character n-grams are extracted from the URL counterparts.

WGI experiments using only the hyperlink information combined (or not) with other web-page information seems to be a promising researching path especially for performance oriented WGI applications such as genre-based focused-crawling where only the URLs are available (Jebari, 2014; Jebari, 2015; Abramson and Aha, 2012; Priyatam et al., 2013) (MSc reference on focused-genre-crawling)

2.3.5 Combination of Features

Instead of using only one type of features, studies in genre recognition tend to combine several sources of information (Lim, 2005). Usually, textual features are considered more important and they are combined with alternative kinds of features . Usually, such combinations increase the effectiveness of the method (Kanaris and Stamatatos, 2009).

An example of combination of textual features from different levels of analysis is reported in (Onan, 2018). The following features are used:

- Most frequent words (function words).

- Character n-grams
- POS n-grams
- Capitalized and lowercase words
- Punctuation marks
- Semantic feature (time and money entities).
- Genre-specific features (n-grams occurring many times within a genre)

In a similar fashion, (Waltinger and Mehler, 2009) combine the following features:

1. Word n-grams
2. Character n-grams
3. POS n-grams
4. Sentence and paragraph length
5. HTML tags
6. HTML attributes
7. Named entities

Other examples of combining different types of features can be seen in Tables 2.1 and 2.2 (Ströbel et al., 2018; Virik, Simko, and Bielikova, 2017). Interestingly, for each feature, the required NLP analysis to extract such measures from documents is also shown. It has to be noted that elaborate types of NLP analysis (e.g. syntactic parsing) introduce a cost concerning the efficiency of the model. In addition, such features are language-dependent.

2.3.6 Domain-specific Genre Representation

Beyond general characteristics that can be extracted from web-pages and be useful in any WGI task, there are domain-specific features related to certain genres and domains that provide a rich representation of their properties.

Blog is a genre with special interest for several research domains and as might be expected it has its own particular characteristics. These features require lexical analysis, morphological analysis, lightweight syntactical analysis, and structural analysis of documents so that they become available. In table 2.2 a rich set of such linguistic properties used for Blog's sub-genres classification are presented in detail. In (Virik, Simko, and Bielikova, 2017) there is a detailed analysis for the correlation of the linguistic features and the Blog's sub-genres. Example of these sub-genres are the following: informative, affecting, reflective, narrative, emotional and rational.

In (Dai, Taneja, and Huang, 2018) the focus is on the News genre. They use a combination of features to recognize the main paradigms of presenting events in news. These features include word unigrams and bigrams, syntactic features like the frequency of syntactic production rules as well as primitive semantic information provided by a pre-defined dictionary (*Linguistic Inquiry and Word Count* (LIWC)). The latter indicates terms that associated with time, motion, and space, important information for quantifying the narrative scheme of the news story. In addition, key events placement features are introduced that attempt to quantify information about specific persons, time, and location of the news story and the point of the document that they occur. In practice, these features calculate the overlap of title with the paragraphs of the document.

Automated genre identification is a subject of interest in the domain of intellectual products (e.g. paintings, music, movies etc). Taxonomies of movies has also a special interest for the technology and entertainment industries. The part of this research related with the current thesis, is when movie genre is induced by textural features such as subtitles and the text description of a video content. Features that are specifically defined for this domain are summarized in Table 2.3. Particularly, BOW, surface and syntactical features are combined. Surface features include content-free and content-specific (the ones related to specific words) information (Lee, 2017). It has been found that not all of these features are so important. The most important of them are the token-type ratio, words per minute, Characters per minute, hapax legomena, dislegomena, short words ratio, ratios of (10, 4, 3, 1)-letter words.

Wikipedia (and in general Wiki sites) is considered as a special genre due to its characteristic, mainly the richness of textual content per page and secondary its informative linguistic register. Also there are several sub-genres of wiki pages which are also characterized as *popular science* web-site and web-documents (e.g. Wikipedia, Nature, New Scientist, Wikinews, etc). There are some domain-specific features that seem to work well for classifying wiki-pages into a sub-genre taxonomy. Table 2.4 shows the set of features used for representing sub-genres of popular science and grouping web-pages with similar properties (Lieungnapar, Todd, and Trakulkasemsuk, 2017).

On the other hand, it is also crucial to study what features used in genre recognition studies remain unaffected by domain variations. This is especially important in genres like News as well as Online reviews. In such cases, it is very important to avoid topic-related information. Ideally, a WGI approach could be trained with samples of a specific topic (e.g., sports) and could be applied to other topics (e.g., politics) without a significant drop in its performance. This is called domain transfer learning (Finn and Kushmerick, 2006). Table 2.5 comprise a topic-neutral set of features (mainly composed of function words and punctuation marks) to achieve this.

2.3.7 Feature Weighting and Selection

Term weighting is an essential issue in text mining applications. The features extracted from web-pages can be represented using a variety of traditional weighting schemes such as Binary representation, Term Frequency (TF), and Term Frequency - Inverted Document Frequency (TF-IDF) (Sharoff, Wu, and Markert, 2010a; Santini, 2007).

The binary scheme is the simplest and according to which each term is represented by a binary value indicating its occurrence or absence in the document. Despite its naivety, very good results were obtained using this scheme in WGI studies kanaris2009learning,sharoff2010web.

TF weighs each term according to its frequency in the document. Several variations of this approach can be found in the literature. For example, the raw frequency of terms can be used. This certainly depends on the length of documents. Another idea is to normalize the raw frequency of a term over text length:

$$TF(t,d) = \frac{f(t,d)}{length(d)} \tag{2.8}$$

where $f(t,d)$ is the raw frequency of term t in document d. Yet another modification is to divide the raw frequency with the maximum frequency of any term in document d.

TF-IDF is a balancing weighting scheme of document terms (e.g., word n-grams, character n-grams, POS n-grams, etc) given a collection of documents. It regulates the significance of the very low and very high frequency terms of the collection. That is, it decreases the value of the very high frequency terms (i.e., function words), and increases the importance of very low frequency terms when they occur in only a few documents. The calculation of a terms IDF in a documents collection is shown in equation 2.9

$$IDF(t) = log\left(\frac{N}{df(t)}\right) \tag{2.9}$$

where N is the number of the documents in the collection and $df(t)$ is the *document frequency* of t, that is the number of distinct documents it occurs.

Although TF-IDF is a popular choice in many text mining studies, the study of (Sugiyanto et al., 2014) demonstrates that it is not the best choice for WGI tasks. On the contrary, they propose a genre-specific weighting scheme, called TF-IGF.

The main idea is that instead of considering a collection of documents, they consider a collection of genres (i.e., each genre is a collection of documents). Then, the terms are weighted by using the frequency of the term within a genre and the *genre frequency* of the term (i.e., the number of different genres it occurs). :

$$TF - IGF(g,t) = f(t,g) \cdot (1 + log\left(\frac{N}{gf(t)}\right) \tag{2.10}$$

where $f(t,g)$ is the frequency of term t in genre g and $gf(t)$ is the genre frequency of t. Since TF-IGF depends on genre, the average value over all genres in a given palette is finally used. The TF-IGF score can be used to select the most informative features that highlight genre-related information and reported results show that it is a better criterion for feature selection in comparison to regular TF-IDF (Sugiyanto et al., 2014).

In (Kanaris and Stamatatos, 2009) a frequency-based method to select the most promising features is described. Initially, the feature set comprises character n-grams of variable length ($n = \{3,4,5$. Then the *LocalMaxs* algorithm is used to find the most prominent n-grams taking into account the frequencies of constituent n-grams of lower order (using a *glue* function). The reported results show that this simple approach is quite effective in WGI tasks.

Another WGI-specific term weighting scheme has been suggested to deal with features obtained from URLs of web-pages (Jebari, 2014). In particular, an approach called *Structure-oriented Weighting Technique* (SWT) first extracts character n-grams from URLs and then each n-gram is weighted according to the following:

$$SWT(t,d) = \sum_{s} w(s) f(t,s,d) \tag{2.11}$$

where $f(t,s,d)$ denotes the raw frequency of n-gram t in section s of document (i.e., URL) d. Namely, this approach assumes that the URL is segmented into fields and each field has its own importance, as follows:

$$w(s) = \begin{cases} \alpha & if \quad s = Domain\ Name \\ \beta & if \quad s = Document path \\ \gamma & if \quad s = Documentname \end{cases} \tag{2.12}$$

Weights $\{\alpha, \beta, \gamma\}$ should be defined empirically using a training corpus (Jebari, 2014).

THERE IS NO REFERENCE FOR THE FOLLOWING WORK (in comments). IN ADDITION THE FORMULAS SEEM PROBLEMATIC AND NOT WELL DEFINED

TABLE 2.1: An example of combining different kinds of features for genre recognition (Ströbel et al., 2018). The NLP analysis required to extract each feature is also shown.

Name	NLP Analysis
Number of Different Words / Sample	Lexical
Correct Type-Token ratio	Lexical
Number of Different Words	Lexical
Root Type-Token ratio	Lexical
Type-Token ratio	Lexical
Lexical Density	Morpho-Syntactic
Mean Length Clause	Morpho-Syntactic
Mean Length Term-Unit	Morpho-Syntactic
Sequence Academic Formula List	Raw text
Lexical Sophistication (ANC)	Raw text
Lexical Sophistication (BNC)	Raw text
Kolmogorov Deflate	Raw text
Morphological Kolmogorov Deflate	Raw text
Syntactic Kolmogorov Deflate	Raw text
Mean Length Sentence	Raw text
Mean Length of Words	Raw text
Words on New Academic Word List	Raw text
Words not on General Service List	Raw text
Clause per Sentence	Syntactic
Clause per Term-Unit	Syntactic
Complex Nominals per Clause	Syntactic
Complex Nominals per Term Unit	Syntactic
Complex Terms Units per Term Unit	Syntactic
Coordinate Phrase per Clause	Syntactic
Coordinate Phrase per Clause	Syntactic
Dependent Clause per Clause	Syntactic
Dependent Clause per Terms Unit	Syntactic
Mean Length of Words (syllables)	Syntactic
Noun Phrase Post-modification (words)	Syntactic
Noun Phrase Pre-modification (words)	Syntactic
Noun Phrase Pre-modification (words)	Syntactic
Term Units per Sentence	Syntactic
Verb Phrase per Term Unit	Syntactic

TABLE 2.2: Blog-specific features and required NLP analysis (Virik, Simko, and Bielikova, 2017).

Name	Description	NLP Analysis
Special characters	Frequency of: @, #, $, %, <WhiteSpace>,&, -, =, +, !, £, ą, [,], /, |	Lexical
Word count	Number of alphanumeric tokens	Lexical
Unique lemmas	Number of unique identified tokens	Lexical
Abbreviations	Ratio of abbreviations to all words	Lexical
Long/short words	Ratio of long (3 or more syllables) to short words	Lexical
Misspelled words	Ratio of misspelled words to all words	Lexical
Nouns	Ratio of nouns to all words	Morphological
Adjectives	Ratio of adjectives to all words	Morphological
Pronouns	Ratio of pronouns to all words	Morphological
Verbs	Ratio of verbs to all words	Morphological
Proper Nouns	Ratio of proper nouns to all words	Morphological
Open/closed words	Ratio of open words (e.g., nouns, adjectives) to open words (e.g., determiners, conjunctions)	Morphological
Functional/content words	Ratio of functional words to content words include nouns, adjectives, numerical, non-modal verbs and adverbs	Morphological
Sequences of functional words	5 or more consecutive functional words with tolerance of one closed word	Morphological
Sentences	Number of sentences	Syntactic
Sentence length	Average sentence length in number of words	Syntactic
Simple/compound sentences	ratio of simple to compound (with two or more clauses) sentences	Syntactic
Sub-sentences	number of simple sentences inside a compound sentence	Syntactic
Dominant tense	Present, future and past	Syntactic
Dominant person	First, second and third	Syntactic
Dominant number	Singular and plural	Syntactic
Links	Ratio of number of Links to number of sections	Structural
Image frequency	Ratio of number of images to number of sections	Structural
Sections	Number of sections	Structural
Section length	Standard deviation words in sections	Structural

TABLE 2.3: Features for video content genre classification (Lee, 2017).

Name	Description	NLP Analysis
Words	Average words per minute	Raw text
Characters	Average characters per minute	Raw text
Word length	Average word length	Raw text
Word n-grams	Frequencies of word n-grams	Raw text
Sentence length	Average sentence length in words	Raw text
Type/token ratio	Ratio of different words to the total number of words	Raw text
Hapax legomena	Ratio of once-occurring words to total words	Raw text
Dis legomena	Ratio of twice-occurring words to total words	Raw text
Short words	Ratio of words with less than 4 characters to total words	Raw text
Long words	Ratio of words with more than 6 characters to total number words	Raw text
Word length	Ratio of words of length of 1-20 to total words	Raw text
Function words	Ratio of function words to total words	Raw text
Descriptive/nominal words	Ratio of adjectives and adverbs to nouns	Syntactic
Personal pronouns	Ratio of personal pronouns to total words	Syntactic
Question words	Ratio of of wh-words to total words	Syntactic
Question marks	Ratio of question marks to total end sentence punctuation	Syntactic
Exclamation marks	Ratio of exclamation marks to total end sentence punctuation	Syntacticc
POS n-grams	Frequencies of POS n-grams	Syntactic

TABLE 2.4: Features used to represent popular science genres (Lieungnapar, Todd, and Trakulkasemsuk, 2017).

Name	Description
Sentence length	Average number of words per sentence
Paragraph length	Average number of sentences per paragraph
Discipline-specific word density	Ratio of specialized vocabulary items to total words
Phrasal verb density	Ratio of phrasal verbs to total verbs
Compound noun density	Ratio of compound nouns to total nouns
Modal verb density	Ratio of modal verbs to total words
Verb density	Ratio of verbs to total words
Adjective density	Ratio of adjectives to total words
Adverb density	Ratio of adverbs to total words
Lexical repetition	Yule's characteristic K
Coordinating conjugation density	Ratio of coordinating conjunctions to total sentences
Content word density	Ratio of content words to total words
Evaluation move density	Ratio of evaluation moves to total sentences
Vocabulary diversity	Probabilities of encountering each word type in 35-50 tokens
Logical connective density	Number of logical connectives per 1000 words
Prepositional phrase density	Number of prepositional phrase per 1000 words
Negation density	Number of negation markers per 1000 words
Pronoun density	Number of pronouns per 1000 words
Flesch reading-ease	Flesh reading-ease index

TABLE 2.5: Topic-neutral features to represent genres (Finn and Kushmerick, 2006).

Type	Features
Surface statistics	Sentence length, Number of words, Words length
Function words	because, been, being, beneath, can, cant, certainly, completely, could, couldnt, did, didnt, do, does, doesnt, doing, dont, done, downstairs, each, early, enormously, entirely, every, extremely, few, fully, furthermore, greatly, had, hadnt, has, hasnt, havent, having, he, her, herself, highly, him, himself, his, how, however, intensely, is, isnt, it, its, itself, large, little, many, may, me, might, mighten, mine, mostly, much, musnt, must, my, nearly, our, perfectly, probably, several, shall, she, should, shouldnt, since, some, strongly, that, their, them, themselves, therefore, these, they, this, thoroughly, those, tonight, totally, us, utterly, very, was, wasnt, we, were, werent, what, whatever, when, whenever, where, wherever, whether, which, whichever, while, who, whoever, whom, whomever, whose, why, will, wont, would, wouldnt, you, your
Punctuation marks	! " $ % ’ ( ) * + - . : ; = ?

2.4 Machine Learning Approaches to Genre Identification

Genre identification of documents is generally viewed as a text categorization task. After defining a feature space to represent documents, a classification algorithm can be applied to a training set in order to learn to distinguish between genres. As already pointed out, the majority of previous work studies consider this to be a closed-set classification task. In addition, most of the existing studies consider a flat genre palette where each genre is independent on the other genres. In the remaining of this section, the machine learning algorithms that have been used to learn the properties of genres are discussed according to the adopted setup of the task.

2.4.1 Closed-set Genre Recognition

The main research volume in this area adopt a closed-set classification framework. Several well-known machine learning algorithms have been used for this task, including SVM, Naive Bayes, Random Forest, Decision Trees, Ensemble-based models (Lim, 2005; Santini, 2007; Kanaris and Stamatatos, 2009; Jebari, 2015; Sharoff, Wu, and Markert, 2010a).

The SVM classifier was tested either in binary or multi-class WGI tasks (Dai, Taneja, and Huang, 2018). It is an algorithm than can easily handle high-dimensional and sparse feature spaces (Joachims, 1997). In (Sharoff, Wu, and Markert, 2010a) analytical experiments using a variety of datasets demonstrated that SVM WGI models could surpass the best reported results in most of the cases combined with character n-gram features. In addition (Virik, Simko, and Bielikova, 2017) compare SVM models with Naive Bayes and k-Nearest Neighbours models on the recognition of Blog sub-genres. The reported results show that SVM obtained higher accuracy results. Recently, an SVM-based approach was tested on the very challenging case of cross-Lingual genre classification (i.e., when the training documents are in one language and the test documents in another language) and obtained very promising results (Nguyen and Rohrbaugh, 2019).

Distance-based approaches in the WGI task include mainly variations of nearest-neighbor classifiers. One particular case is based on ranked feature distributions distances (Waltinger and Mehler, 2009). The features of the samples of a class are ranked in descending order according to their TF or TF-IDF values. In order to measure the distance of a new web-page from the classes, the features of the new web-page are also ranked and then the difference in rankings indicate the most similar class. That is the TF or TF-IDF value of features is not important anymore since only the ranking of features is considered. Moreover, when a feature is not present in either the new web-page or a class, then a predefined *Max* value is assigned. The total *ranking distance* between a web-page *d* and a class *g* is calculated as follows:

$$d(d,g,t) = \begin{cases} |r_d(t) - r_g(t)|, t \in d \wedge t \in g \\ Max, t \notin d \vee t \notin g \end{cases} \tag{2.13}$$

$$rd(d,g) = \sum_t d(d,g,t) \tag{2.14}$$

The new web-page is then classified to the nearest class. The accuracy of this method has been reported to surpass that of SVM using the same features (Waltinger and Mehler, 2009).

Following the impressive performance obtained in classification tasks involving natural language texts, deep learning algorithms have also been tested in WGI tasks (Ströbel et al., 2018). A *recurrent neural network* comprising 200 gated recurrent unit cells in the hidden layer. On top of that, a fully-connected layer assigns documents to classes using a Softmax decision function. Very promising results are reported for this deep learning model in closed-set WGI tasks.

In another recent study, a variety of deep learning algorithms are compared with traditional methods and the latter seem to be more accurate in genre identification tasks (Worsham and Kalita, 2018). In more detail, a convolutional neural network, a long short-term memory network and a hierarchical attention network have been applied to recognition of literary genres. However, they were outperformed by relatively simple models based on traditional machine learning algorithms. In addition, deep learning methods considerably increase the training time cost and require special hardware infrastructure to handle long texts.

Instead of learning a simple model, ensemble methods attempt to extract several base models and then combine them. One main direction is to use well-known ensemble learning methods such as AdaBoost, Bagging and Random Forests (Sugiyanto et al., 2014; Onan, 2018; Worsham and Kalita, 2018). This approach can easily handle high-dimensional representations and heterogeneous features.

Although, the traditional bag-of-words approach had better result with XABoost or other techniques been tested for over a decade on genre identification or/and particularly on WGI, distributional feature models are early showing their advantages over the TF-IDF (or TF alone) models[REF].

Another idea is to build a separate model for each web-page modality. For example, an ensemble algorithm called *Multiple Classifier Combination* (MCC) is presented in (Zhu et al., 2016). Particularly, the main idea is use information from a web-pages to be classified to a given genre palette as well as information from a set of neighbouring web-pages (i.e., that are near the specific web-page in the graph formed by hyperlinks between pages). The MCC algorithm builds a set of SVM classifiers each trained using a particular set of features. Then a decision matrix is formed including all predictions of base SVM classifiers:

$$DP(p) = \begin{pmatrix} d_{11}(p) & \cdots & d_{1|G|}(p) \\ d_{21}(p) & \cdots & d_{2|G|}(p) \\ & \vdots & \\ d_{N1}(p) & \cdots & d_{N|G|}(p) \end{pmatrix} \tag{2.15}$$

where d_{ij} is the membership degree given by classifier i to genre j, N is the number of base classifiers, and $|G|$ is the number of genres. Then, the final decision is taken by applying simple methods to combine these predictions columnwise, such as the min, max or average rules.

Another *late fusion* ensemble is proposed in (Finn and Kushmerick, 2006). Again, the idea is to build homogeneous base models each trained only on a specific feature subset. In the testing phase the majority voting us a common strategy. Particularly in their study they learn C4.5 decision trees for different web-page modalities (i.e., BOW, POS, text statistics features) and then build a *Multi View Ensemble* that combines the predictions of the modality-specific models. It is important to note that in the training phase *Active Learning* was used. This is a sample selection strategy where an an evaluating process was indicating which sample was better to be used for the specific C4.5 learner, for a given feature set. The late fusion ensemble with the active learning strategy obtained promising results including the domain transfer scenario.

2.4.2 Open-set Classification

Most previous studies in WGI consider the case where all web pages should belong to a predefined taxonomy of genres (Lim, 2005; Santini, 2007; Kanaris and Stamatatos, 2009; Jebari, 2014). This corresponds to the closed world assumption. However, this naïve assumption is not appropriate for most applications related to WGI since it is not possible to construct a universal genre palette that covers at least a great extend of the Web. Web-pages that do not belong to any of the predefined genres are considered noise and also include web-pages where multiple genres co-exist (Santini, 2011; Levering, Cutler, and Yu, 2008).

Noise in WGI can be categorized into structured noise and unstructured noise. The former assumes that there is no information about the composition of noise (i.e., a random collection of web-pages not belonging to the known genres) (Santini, 2011). The latter assumes that noise is composed by several unknown genres (i.e., for which there are no training examples). However, it is highly unlikely that such a collection represents the real distribution of pages on the web.

The effect of noise in WGI was first studied in (Shepherd, Watters, and Kennedy, 2004; Kennedy and Shepherd, 2005) where predefined genres were personal, organizational, and corporate home pages while noise consisted of non-home pages. However, the distribution of pages into these four categories was practically balanced, hence it was not realistic. In another study, a clustering framework is used where