



UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

TagTheWeb: USING WIKIPEDIA CATEGORIES TO AUTOMATICALLY
CATEGORIZE TEXT-BASED RESOURCES ON THE WEB

Jerry Fernandes Medeiros

Orientadores

Bernardo Pereira Nunes

Sean Wolfgang Matsui Siqueira

RIO DE JANEIRO, RJ - BRASIL

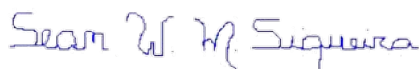
SETEMBRO DE 2018

TagTheWeb: USING WIKIPEDIA CATEGORIES TO AUTOMATICALLY
CATEGORIZE TEXT-BASED RESOURCES ON THE WEB.

Jerry Fernandes Medeiros

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA
OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-
GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO
DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA
ABAIXO ASSINADA.

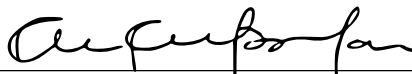
Aprovada por:



Sean Wolfgang Matsui Siqueira, D.Sc. – UNIRIO



Bernardo Pereira Nunes, D.Sc. - UNIRIO



Ana Cristina Bicharra Garcia, D.Sc. - UNIRIO



Terhi Nurmikko-Fuller, PhD - ANU (Austrália)

Catálogo informatizada pelo(a) autor(a)

M488 Medeiros, Jerry Fernandes
Using Common Sense Knowledge to Automatically
Classify Resources on the Web / Jerry Fernandes
Medeiros. -- Rio de Janeiro, 2017.
100

Orientador: Bernardo Pereira Nunes.
Coorientador: Sean Wolfgang Matsui Siqueira.
Dissertação (Mestrado) - Universidade Federal do
Estado do Rio de Janeiro, Programa de Pós-Graduação
em Informática, 2017.

1. Text Categorization. I. Nunes, Bernardo
Pereira, orient. II. Siqueira, Sean Wolfgang
Matsui, coorient. III. Título.

“When we are no longer able to
change a situation, we are
challenged to change ourselves.”
Viktor Frankl

I dedicate this thesis to my mother
and father;

Medeiros, Jerry Fernandes. **TagTheWeb: USING WIKIPEDIA CATEGORIES TO AUTOMATICALLY CATEGORIZE TEXT-BASED RESOURCES ON THE WEB.** UNIRIO, 2018. 91 pages. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

A identificação de tópicos associados a um conjunto de documentos é uma tarefa comum para muitas aplicações e pode ser usada para melhorar diversas tarefas envolvendo documentos na Web, tais como a busca, recuperação da informação, recomendação, armazenamento e agrupamento. Devido à quantidade significativa de informações produzidas e disponibilizadas hoje na Web, torna-se humanamente impossível organizar, analisar e extrair o conhecimento incorporado nesses documentos. Consequentemente, mecanismos para realizar tarefas como remover ou pelo menos diminuir a necessidade de intervenção humana ganharam importância nas últimas décadas. Uma das possíveis soluções para lidar com o desafio de organizar e recuperar documentos é usar classificação automatizadas de informações. Nesta pesquisa, propõe-se um método de classificação genérico para categorizar automaticamente conteúdo baseado em texto na Web de acordo com o conhecimento coletivo dos colaboradores da Wikipédia, por meio da relação semântica entre os nós do Gráfico de Categoria da Wikipédia. A abordagem é baseada em três etapas: extrair entidades nomeadas do texto, extrair categorias associadas a entidades nomeadas e, finalmente, representar e classificar o documento. Para validar o método aplicado, foram realizados experimentos computacionais e um estudo envolvendo usuários de uma plataforma de crowdsourcing. Os resultados mostram que a abordagem aplicada é capaz de categorizar corretamente a maioria dos documentos de uma maneira que os usuários reais possam entender, sem o esforço dos especialistas em domínio.

palavras-chave: Classificação de texto, Wikipédia, Categorias, Grafo de Categorias

ABSTRACT

Identifying topics associated with a set of documents is a common task for many applications and can be used to improve various tasks involving documents on the Web, such as search, retrieval, recommendation, and clustering. Due to the significant amount of information produced and made available today, it becomes humanly impossible to organize, analyze, and extract the knowledge embedded. Consequently, mechanisms to accomplish such tasks as removing or at least diminishing the need for human intervention has gained importance in the last decades. One of the potential solutions for dealing with the challenge of organizing and retrieving documents is to use automated classification and categorization of Web information. In this research, a generic classification method to automatically categorize any text-based content on the Web according to the collective knowledge of Wikipedia contributors, through the semantic relation between nodes of the Wikipedia Category Graph, is proposed. The approach is based on three steps: extracting named entities from text, extracting categories associated with named entities, and finally representing and classifying the document. Computational experiments and a study involving users of a crowd-sourcing platform were used to validate the method. The results show that this approach can be used to correctly categorize most documents in a way that real users can understand, without the effort and input of domain experts.

Keywords: Text Classification, Wikipedia, Categories, Category Graph

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 1 |
| 1.1 | Contextualization | 1 |
| 1.2 | Motivation | 4 |
| 1.3 | Problem Statement | 5 |
| 1.4 | Goals of this Thesis | 5 |
| 1.5 | Project Overview | 6 |
| 1.6 | Main Contributions | 7 |
| 1.6.1 | Scientific Contributions | 7 |
| 1.6.2 | Technical Contributions | 7 |
| 1.7 | Thesis Outline | 8 |
| 2 | Collective Knowledge and Wikipedia | 10 |
| 2.1 | Collective Knowledge Construction | 10 |
| 2.2 | Wikipedia | 11 |
| 2.2.1 | Wikipedia Structure | 13 |
| 2.2.1.1 | Titles and Wikilinks | 13 |
| 2.2.1.2 | Infoboxes | 15 |
| 2.2.1.3 | Categories | 16 |
| 2.2.2 | The Wikipedia Category Graph (WCG) | 18 |
| 3 | The Wikipedia Category Graph | 20 |
| 3.1 | The Category Graph | 21 |

| | | |
|----------|---|-----------|
| 3.2 | Small-world networks (SW) | 22 |
| 3.2.1 | Sparsity and Connectedness | 23 |
| 3.2.2 | Path lengths | 24 |
| 3.2.3 | Clustering Coefficient | 25 |
| 3.2.4 | Degree Distribution | 26 |
| 3.2.5 | Empirical demonstration of small-worldness | 28 |
| 3.3 | Final Consideration | 29 |
| 4 | Fundamental Concepts | 31 |
| 4.1 | Information Retrieval | 31 |
| 4.1.1 | Documents Representation | 33 |
| 4.1.2 | Named Entity Recognition | 35 |
| 4.2 | Automatic Text Classification | 36 |
| 5 | Methods | 39 |
| 5.1 | Approach | 39 |
| 5.1.1 | Text Annotation | 40 |
| 5.1.2 | Categories Extraction | 42 |
| 5.1.3 | Representation of Document | 44 |
| 5.2 | Final Consideration | 47 |
| 6 | Experiments, Results, and Discussion | 49 |
| 6.1 | Proof of Concept - Q&A Communities | 49 |
| 6.1.1 | Resources and Methods | 50 |
| 6.1.2 | Results and Discussion | 50 |
| 6.2 | Crowdsourcing study | 54 |
| 6.2.1 | Experimental Design | 55 |
| 6.2.2 | Quality Control | 56 |
| 6.2.3 | Results and Discussion | 58 |
| 6.2.3.1 | Elite workers vs Regular workers | 59 |

| | | |
|-------------------|---|-----------|
| 6.2.3.2 | Human Judgment Analysis | 59 |
| 7 | Related Works | 66 |
| 8 | Final Remarks, Limitations and Future Works | 70 |
| 8.1 | Final Remarks | 70 |
| 8.2 | Contributions | 71 |
| 8.3 | Limitations | 72 |
| 8.4 | Future Works | 73 |
| Appendix A | Representing the Underlying Structure of the WCG | 75 |
| Appendix B | Percentage Distribution of Categories | 78 |
| Appendix C | Crowdsourcing Experiment Details | 82 |
| Appendix D | Wikipedia Category Graph - Nodes Dataset | 84 |

List of Figures

| | | |
|-----|--|----|
| 2.1 | Disambiguation page for the term “Mercury” | 15 |
| 2.2 | Example Infobox of the page related to the god Mercury with the default properties for infoboxes about deities. | 16 |
| 2.3 | Example of an induced graph showing the supercategories of “Semantics” in the Wikipedia Category Graph. | 17 |
| 2.4 | Example of an induced graph showing the categories and relationships for the Entity Apple towards Main topics | 18 |
| 3.1 | Simplified example of the underlying structure of the Wikipedia Category Graph (WCG) | 22 |
| 3.2 | A tree-structured hierarchy | 23 |
| 3.3 | Scale-free small-world graph | 23 |
| 3.4 | An arbitrary, unstructured random graph | 23 |
| 3.5 | Structures of semantic networks adapted from Steyvers and Tenenbaum [66] | 23 |
| 3.6 | Shortest path length distribution on the WCG | 25 |
| 3.7 | Example of how the clustering coefficient of a node is calculated based on the probability of the neighbors of the node i are also neighbors among themselves | 26 |
| 3.8 | Indegree and outdegree distributions | 28 |
| 4.1 | Example of the Named Entity Recognition (NER) task on a text extracted from a social network. | 36 |
| 5.1 | An induced graph containing all shortest paths from the categories found and described in table 5.2 and “Main topic classifications” (on the right of the graph) | 46 |

| | | |
|-----|--|----|
| 5.2 | Final classification of the running example according to our method | 47 |
| 6.1 | The number of shortest paths through the proposed method. The x-axis shows the number of paths found for each top-level category (displayed on the y-axis) | 53 |
| 6.2 | Percentage distribution of answers given by crowd contributors for each one of the ten communities evaluated | 62 |
| B.1 | Percentage distribution of categories for each of the communities evaluated according to the proposed method. | 81 |
| C.1 | Example of task delivered to contributors in the dataset Biology | 82 |
| C.2 | Example of feedback given by the consulting service provided by Crowd-Flower platform | 83 |

List of Tables

| | | |
|-----|---|----|
| 3.1 | Power-law parameters of the WCG | 27 |
| 3.2 | Empirical demonstration of small-worldness of the WCG. | 29 |
| 5.1 | Entities extracted from the text and their respective links to DBpedia concepts. | 42 |
| 5.2 | List of all entities extracted from the example text and the categories associated to them | 43 |
| 5.3 | Example of a path for each top-level category that contributed to the document representation | 47 |
| 6.1 | Distribution of post type in the stack exchange datasets along with the average text length | 51 |
| 6.2 | Overall setup for the experiment with crowd workers | 59 |
| 6.3 | Values of precision, recall and F-measure when comparing our classification and the judgments made by workers in the crowdsourcing study | 64 |
| A.1 | Description of fields in INSERT statements for the table categorylinks. . . . | 76 |
| D.1 | A sample including the 50 first rows of the dataset generated from the Wikipedia Category Graph. Each row represent one node of the graph along with its Degree, Clustering Coefficient, and Centrality information | 84 |

Glossary

BOW Bag of Words. 34, 35, 67, 68

FG Features Generation. 35

IE Information Extraction. 40

IR Information Retrieval. 1, 30–33, 38, 49, 57, 69, 71, 77

KNN K-nearest Neighbors. 38

LOD Linked Open Data cloud. 6

NER Named Entity Recognition. vii, 35, 36, 40, 41

NLP Natural Language Processing. 20, 30, 71

ODP Open Directory Project. 1

OKG Open Knowledge Graph. 40

RDF Resource Description Framework. 42

SPARQL SPARQL Protocol and RDF Query Language. 42

SVM Support Vector Machines. 38

VSM Vector Space Model. 33, 34, 44

WCG Wikipedia Category Graph. vii, ix, 18–30, 39, 45, 47, 50, 54, 71, 72, 75

WWW World Wide Web. 23, 32

1. Introduction

1.1 Contextualization

The organization of information has been a concern of human beings since the beginning of the first civilizations, about 4,000 years ago [3]. At that time, accounting records, government directives, contracts, and court sentences were kept and organized into clay tablets. Over the years, these tablets have been replaced by paper, and the number of documents increased considerably. Hence the activity of locating them had become a significant challenge for the organization of information.

An example of this is in the classification of books in a library. Librarians ordinarily use classification systems to organize the books on their shelves, clustering together those that are on the same topic. The topics themselves are usually divided into increasingly specific subcategories, forming a hierarchical classification.

In the early 1990s, the Web appeared. It represented a distributed hypermedia system that enabled users to search for information in a wide variety of areas of knowledge. The large volume of Web documents and the inability to perform extensive editorial control in this system have contributed to the emerging importance of the organization of Web documents, a substantial challenge facing Information Retrieval (IR), a research area that deals with the problem of representing, organizing and storing information for the user to access them using the computer [3]. Web directories such as the Open Directory Project (ODP), Yahoo! Directory and Google Directory are applications that try to organize Web

documents into a hierarchy of topics to make it easier to navigate and retrieve them.

The expansion and maintenance of these directories have been done manually by publishers who analyze the content of Web documents and classify them on particular topics. These manual classification were however ineffective, due mainly to the number of documents published on the Web, and all of them have been discontinued at some point.

If on the one hand the expansion of the Web has represented a challenge for experts aiming to classify the raising number of documents, then on the other, it has brought to the fore a significant social impact by enabling users to participate in the construction and organization of information. Folksonomies, for instance, are collaborative attempts to categorize items of some type, with the aim of helping users in their searches [54].

In this context of the collaborative construction of knowledge on the Web, Wikipedia is the most substantial encyclopedia freely available. It has been developed and curated by a large number of users over a period of time, and contains information covering a very broad range of topics currently found on the Web. Wikipedia is itself organized with a folksonomy, one that takes the form of a category hierarchy: to each Wikipedia article there are one or more categories, which are themselves structured as a collaborative hierarchical structure. Therefore, Wikipedia can be considered as a knowledge graph with an explicit, human-authored form of classification.

As in many classification methods, such as Dewey Decimal Classification[50] and Library of Congress Classification[8], an article in Wikipedia can belong to one or more top categories which in some sense represent the topics it covers. Within Wikipedia, the primary purpose of this classification is to facilitate the search for relevant information.

The Wikipedia Categorization scheme is a thesaurus collaboratively constructed and used for indexing the content of Wikipedia pages [72]. Hence, it can be said that it represents a common or shared understanding of Wikipedia's content. It is a classification made by the user community, rather than one elicited from experts for laypersons. The richness of this type of information, which enables several tasks performed by users on the

Web (such as search, information retrieval, recommendation, and clusterization), is also noteworthy.

Categorization plays a crucial role in the future of information search services, and many positive categorization approaches involve the integration of knowledge from Wikipedia. This explains the widespread use of Wikipedia's article contents and category hierarchy to generate semantic resources that enhance performance on text classification and keyword extraction among other applications [20].

In this context, the primary purpose of this thesis is to create a general-purpose classification method based on the Wikipedia Categorization scheme to categorize text-based content on the Web, for instance, scientific articles, web pages or even posts on social media. Although the API is versatile enough to generate a classification for any textual input, the validity of the method in the scope of this thesis was tested only in a small extent, with texts extracted from question and answer communities. The method relies on feature extraction from the collective knowledge of Wikipedia contributors, rather than on a traditional classification system created by domain experts. That is, regular users are more likely to successfully access and retrieve information from a Web created by people for people, than from one created solely by experts.

At its current stage, the proposed method can classify content in English and can be accessed at <http://www.TagTheWeb.com.br>.

The Wikipedia categorization scheme was chosen for four main reasons:

- Wikipedia is the largest online encyclopedia and is constructed and maintained by contributors from all over the world.
- Wikipedia contains an categorization scheme, curated by humans, where all articles are placed within categories that describe their content, and these categories are semantically related to other categories in a rich and meaningful network.
- Wikipedia categories are words or phrases in natural language, making them easy

for regular users of information retrieval systems to understand and interact with.

- The content on Wikipedia is dynamic, allowing an adaptive and evolving classification method. Unlike other encyclopedias, Wikipedia is often updated in real time.

1.2 Motivation

With the ubiquitous Internet and the rapid growth of the Web, accessing the vast amount of digital text remains a challenge for users. One of the potential solutions is to use automated classification and categorization of Web information [14].

The amount of data available in digital format on the worldwide Web has increased steadily. According to estimates made in 2014, from 2013 to 2020 the digital universe will increase from 4.4 trillion gigabytes to 44 trillion gigabytes [70]. Much of the data in the digital universe is in textual formats, such as emails, reports, newsletters, articles, and Web page content. Also, with the advent of the Web 2.0, textual data has been used as a means to disseminate information, whether by postings on social networks, wikis or blogs [13] [53].

Due to the significant amount of textual information produced and made available today, it becomes humanly impossible to organize, analyze, and extract knowledge embedded in textual information. Consequently, mechanisms to accomplish such tasks as removing or at least reducing the need for human intervention have gained importance in recent decades [11] [5] [1].

The possibility of investigating a method for automatic text classification, capable of assigning categories that are understandable to humans and take into account the collective knowledge encoded into Wikipedia rather than an expert's effort, is what motivates the development of this work. Text classification allows users to find desired information faster by searching relevant categories only (rather than the entire information space), and helps users to develop conceptual views of digital documents. Since the information exists in unstructured form, categorization can allow users to make the most use of texts.

1.3 Problem Statement

According to Sebastiani [61], although the first efforts to automate the classification of digital documents were made in the 1960s, a semiautomatic technique based on knowledge engineering was used for document classification until the 1980s.

Whilst a semiautomatic approach can be precise, it has a significant limitation concerning the acquisition of knowledge for the construction of the classifier. This limitation is primarily the need to have at least two human experts involved in the process: a domain expert, with the ability to classify documents in the predefined set of classes; and a knowledge engineer, able to encode the classification in a programming language as a set of rules. This approach is inflexible because each and any iterative stage of changes or new developments in the classification system necessitates the involvement of the two experts to adjust the rules and the classifier.

1.4 Goals of this Thesis

The primary goal of this research is to study the viability of taking advantage of this collective body of knowledge to automatically categorize web-based content according to Wikipedia contributors. The primary objective can be broken down into the following specific goals:

- i Perform a graph-theoretic analysis of the Wikipedia Category Graph, to describe the topology of this structure and to identify the challenges and potentials of employing it for extracting features from Web documents;
- ii Propose an approach for the extraction of features in text-based resources based on the intelligence embedded in the structure of Wikipedia categories
- iii Suggest a method for representing documents, which is based on Wikipedia categories, and allows for automatic classification;

- iv Design, execute and analyze the results of an experimental study involving crowd-sourcing, to verify whether humans recognize the classification generated by the proposed method.

1.5 Project Overview

Automatic text classification is a process where a category or a set of categories are assigned to a textual resource, based on specific criteria.

There are several methods for performing automatic text categorization. This project focuses on categorizing text based on the named entities found in the text and its relation to a set of predefined categories.

A processing chain to generate a generic categorization consists of three steps:

(i) Text Annotation;

Automatically extract structured information from unstructured text and link it to an external knowledge base in the Linked Open Data cloud (LOD). For this thesis, DBpedia was used because it is based on information extracted from Wikipedia.

(ii) Categories Extraction;

In this step, the entity relationships are traversed to find a more general representation of the entity: their categories. All categories associated with the entities identified in the text are extracted and indexed.

(iii) Document Representation.

The set of all Wikipedia categories cannot be directly used as a feature for categorization, because different texts will have a different set of categories, making it impossible to categorize and compare them. To reduce this dimensionality, the applied approach consists of navigating the Category Graph from each category extracted in the previous step towards the top of the graph by all the shortest paths between the

category and the main topics.

Based on the influence of each main topic category on the resource, a document representation of the calculated categorization was generated as a multidimensional vector.

1.6 Main Contributions

This thesis has two primary categories of outcomes: i) Scientific Contributions and ii) Technical Contributions.

1.6.1 Scientific Contributions

- The proposal of a method for the extraction of features and representation of text-based resources, based on the categorization scheme of Wikipedia.
- The results of the experimental crowd-sourcing study indicating a positive correlation between the classification generated by the proposed method and the understanding of people about the content of the documents evaluated.
- The results of an updated analysis of the Wikipedia Category Graph, which indicates that like other networks used for natural language processing problems, it is also a small-world and scale-free network.

1.6.2 Technical Contributions

- TagTheWeb¹, a public, documented² and open-source API capable of receiving any textual resource and processing each of the three phases described in the proposed approach (see 5.1).
- The Wikipedia Category Graph snapshot from October of 2016 filtered and repre-

¹<http://www.tagtheweb.com.br>

²<http://documenter.getpostman.com/view/1071275/tagtheweb/77bc7K>

sented in Neo4J³ and graph-tools⁴.

- A dataset⁵ containing all nodes of the Wikipedia Category Graph and the measures of centrality, in-degree, out-degree, clustering coefficient, and PageRank. An example with the first 50 categories can be seen in Appendix D.

Part of this research has already been published: *TagTheWeb: Using Wikipedia Categories to Automatically Categorize Resources on the Web* [43].

1.7 Thesis Outline

This thesis is organized into eight chapters, this Introduction being the first of them. The other chapters are organized as follows, and describe, respectively:

- **Chapter 2:** The main features and the organization of Wikipedia, the primary source of information.
- **Chapter 3:** The graph-theoretic analysis carried out on the Wikipedia Category Graph.
- **Chapter 4:** The fundamental concepts needed to understand the method described in this thesis.
- **Chapter 5:** The steps of the proposed method illustrated by a running example.
- **Chapter 6:** The evaluation methods employed, and the results of computational and crowd-sourcing experiments.
- **Chapter 7:** The closest related works that served as inspiration for the development of this research.

³<http://www.neo4j.com>

⁴<http://www.graph-tool.skewed.de>

⁵<http://www.github.com/jerrylewisbh/TagTheWeb>

- **Chapter 8:** The conclusions extracted from the experiments, the contributions of the research in a general context, its limitations, and perspectives of future work.
- **Appendix A** details how the information was extracted from Wikipedia and represented as a directed graph.
- **Appendix B** presents the percentage distribution of categories along ten stock exchanged communities used in the experiment described in section 6.1.
- **Appendix C** contains details about the crowd-sourcing experiment described in section 6.2.
- **Appendix D** Shows a sample of the dataset containing all nodes of the Wikipedia Category Graph and the measures of centrality, indegree, outdegree, clustering coefficient, and PageRank.

2. Collective Knowledge and Wikipedia

This chapter briefly describes the concept of Collective Construction of Knowledge as it is essential for understanding the motivation behind choosing an approach that uses the knowledge of the contributors of Wikipedia, rather than experts. A detailed description of Wikipedia and its features (central to understanding the organization of this body of knowledge), as well as the possibilities and challenges that emerge from decoding its underlying structure are also presented.

2.1 Collective Knowledge Construction

Pierre Lévy [36], a French philosopher who specializes in the understanding of the cultural and cognitive implications of digital technologies and the phenomenon of collective human intelligence, argues that knowledge is in humanity, and every individual can offer knowledge.

Cyberspace allows individuals to remain interconnected regardless of their geographic location. It deterritorializes knowledge and supports the development of collective intelligence. An essential factor in the efficient mobilization of competences is the identification and understanding of the capabilities of the subjects. Lévy's project of collective intelligence is not only linked to cognition. It is also a global project that presumes practical actions intended to mobilize the competences of individuals to provide mutual recognition and enrichment of those who are involved in this proposal [36].

Lévy [37] defines collective intelligence as a new sustainable way of thinking through social connections that become viable through the use of a network of people on the Web. This collective intelligence is distributed and coordinated in real time, which results in an efficient mobilization of skills, and the cyberspace favors its development.

2.2 Wikipedia

Wikipedia is the most substantial encyclopedia freely available on the Web. It has been developed and curated by a large number of users over time and represents the result of a process of collective construction about facts, people and the broadest type of topics currently found on the Web.

Wikipedia content is available in around 300¹ active languages. The English version has more than 5.4M articles, written and edited by a total of some 30 million registered editors, of whom roughly 120,000 are currently active. In the last ten years, there has been a consistent average of 30 million edits per year, including both the creation of new articles, and the development of existing ones.

This online encyclopedia was created in January 2001 as an improvement of Nupedia, a similarly free encyclopedia, but one written only by specialists with rigid evaluation criteria. It had low adherence, and was suspended in 2003. Both the Wikipedia and Nupedia projects were initiatives by Jimmy Wales and Larry Sanger. At the beginning of 2008, Wikipedia exceeded 8 million entries in 253 languages. It proceeded to double the number of entries at an annual rate for the following few years, making it currently the fifth most accessed site in the Web².

Its success among users and its dissemination as a source of reference do not lie in the fact that Wikipedia is on the Internet, since there are other alternatives available online. What differs is the possibility of participation, collaboration, and collective construction.

¹https://en.wikipedia.org/wiki/List_of_Wikipedias

²<http://www.alexa.com>

The Wiki system allows not only the gathering of data, but also the collective generation of new knowledge across different subjects. In this regard, Wikipedia is not merely a tool for indexing and formatting, but a space for debating and synthesizing texts. The contributors are not just “librarians”, but authors, in the strictest sense of the word. Wikipedia is more than a source of information; it is also an invitation to collaborative knowledge construction. While the use of a conventional encyclopedia risks querying for information that has already become dated by the time of its publication, and whose volumes rest immutable on the shelf, Wikipedia opens its pages to the present and the ongoing debate over available writings. Each participant contributes by offering questions for discussion. Through these mutual exchanges, the text of the entries is discussed and improved. When the inclusion of dubious information compromises text, new discussions and corrections can be initiated. Some authors have shown that vandalism and inaccuracies in Wikipedia are often reverted within a matter of minutes [31] [71].

A study by Wilkinson & Huberman [76] indicated that the popularity of the project and the reliability of many of the texts is a result of the intense participation of registered users. The thousands of volunteers who contribute to the project make the site an environment of intense social interaction, in which each user fulfills specialized functions, according to their interest, availability and (eventually) bureaucratic role.

Seeking to increase the reliability of content built in a collective and collaborative environment, Wikipedia has created a rigid organizational structure. Note that, as Tapscott and Williams [69] affirm, collaborative production mixes elements of hierarchy and self-organization and is based on meritocratic principles of organization.

The editing community enforces specific codified rules designed to ensure accuracy and prevent bias. A study comparing the precision of various scientific subjects in Wikipedia and Encyclopaedia Britannica found that while errors were not infrequent, they occurred at similar rates between the two [21]. In particular, Wikipedia science articles contained an average of four mistakes, while Encyclopaedia Britannica ones included only three. The latter currently has about 65,000 articles, while the English Wikipedia

has approximately 5.4 million (totaling 1.8 billion words). Wikipedia is a free online encyclopedia where all readers can update content by including and editing articles. Instead of following a peer review process by experts, revisions and enhancements are contributed by readers. In [44], the sophisticated techniques for extracting knowledge from different perspectives developed by researchers is demonstrated:

- Wikipedia as an encyclopaedia;
- Wikipedia as a corpus;
- Wikipedia as a thesaurus;
- Wikipedia as a database;
- Wikipedia as an ontology; and,
- Wikipedia as a graph.

Although Wikipedia texts are written in natural language, some structured resources are available for organizing articles into categories, for connecting different articles, and for presenting the relevant properties of the topic described in the article.

2.2.1 Wikipedia Structure

Wikipedia has different types of elements in its structure. In this section, we describe the main features that make up the organization of Wikipedia and that are relevant for the automatic extraction of knowledge.

2.2.1.1 Titles and Wikilinks

Each Wikipedia article has a name, which is the most common form of identification of the concept or entity described in the article. People, organizations, places, events, and species of living beings are common classes described on Wikipedia. The titles are unique identifiers within the set of Wikipedia articles for a language.

The guarantee of the uniqueness of the title makes it possible to reference an article through its title. Wikipedia explores this possibility through internal links (Wikilinks). Wikilinks are references that enable navigation between articles in a network of internal links built by the publishers of the articles.

Wikipedia recommends that editors link only the first occurrence of a reference to another article through a Wikilink. It is also possible to separate the link itself from the term it refers to, thus creating an arbitrary alternative text for the link. This process is often used for homonyms and abbreviations and can be applied by adding a pipe “|” divider followed by the alternative name. The article comes before the divider and the text that is displayed and placed after it. For example, the formatting of the link `[[List of Presidents of the United States|44th President of the United States]]` ensures that the final article will display only the 44th President of the United States in the text, with a clickable link leading to the List of Presidents of the United States article on Wikipedia.

Homonyms (single words that represent different concepts or entities) stand to violate this restriction of uniqueness for Wikipedia titles. In these cases, the article that defines the most known concept remains with the simple name, and the other titles must have a suffix for disambiguation. The suffix of disambiguation must present a detail that makes the differentiation of one article from the others possible. It is suggested that editors create a specific disambiguation page that lists the different articles related to a specific homonym with internal links to their contents. When it is not possible to determine which of the concepts is best known, the disambiguation page has the simple title, and all other pages have the suffix for disambiguation.

An example of disambiguation on Wikipedia in English is the concept “Mercury” (see figure 2.1, which can mean:

- a metallic chemical element with the symbol “Hg”;
- a Roman god; and,
- the first planet from the Sun.

Since it is not possible to determine the most known entity, the disambiguation page has the title “Mercury”³; the planet has the title “Mercury (planet)”⁴; the god has the title “Mercury (mythology)”⁵; and the element is named “Mercury (element)”⁶.

Mercury

From Wikipedia, the free encyclopedia

Mercury usually refers to:

- [Mercury \(element\)](#), a metallic chemical element
- [Mercury \(mythology\)](#), a Roman god
- [Mercury \(planet\)](#), first planet from the Sun

Figure 2.1: Disambiguation page for the term “Mercury”.

Another variant of Wikilinks is the redirect, employed when different textual forms refer to a single concept or entity. This situation would cause a conflict with the restriction of the uniqueness of titles, imposing the repetition of the content but with different titles.

The redirect pages contain only text in the form of a directive without gender, number, or case. The central purpose is to find a single article for equivalent terms. For example, if the user searches for “apples” (plural) the redirect page will refer them to the “apple” (singular) article.

Redirections also occur with people’s names, when they are known both by their full name and by part of their name or surname. It is the case of the English writer J. R. R. Tolkien, well-known only by his last name, Tolkien.

2.2.1.2 Infoboxes

According to Wikipedia documentation⁷, infoboxes are fixed-format tables that summarize relevant aspects of an article. It is an optional feature, but they present common attributes between different subjects. Wikipedia recommends the use of a predefined infobox template, as they already have known suggested attributes.

³<https://en.wikipedia.org/wiki/Mercury>

⁴[https://en.wikipedia.org/wiki/Mercury_\(planet\)](https://en.wikipedia.org/wiki/Mercury_(planet))

⁵[https://en.wikipedia.org/wiki/Mercury_\(mythology\)](https://en.wikipedia.org/wiki/Mercury_(mythology))

⁶[https://en.wikipedia.org/wiki/Mercury_\(element\)](https://en.wikipedia.org/wiki/Mercury_(element))

⁷<https://en.wikipedia.org/wiki/Help:Infobox>

When editors use predefined infoboxes in an article, Wikipedia displays the table with special formatting that enriches the visual aspect of the box. They are also used as metadata by projects such as DBpedia. Figure 2.2 shows the infobox for the article on Mercury (a god in Roman religion and mythology).

| Mercury | |
|--|---|
| God of financial gain, commerce, messages/communication, travelers, boundaries, luck, trickery, merchants, thieves | |
|  | |
| Consecration relief with the god Mercury (right). A man is offering a goat at an altar | |
| Symbol | Caduceus, winged sandals, winged hat, tortoise, ram and rooster |
| Personal information | |
| Consort | Larunda |
| Children | Lares |
| Parents | Maia and Jupiter |
| Greek equivalent | Hermes |

```

{{Infobox deity
| type           =
| name         =
| member_of     =
| image         =
| alt         =
| image_size    =
| caption     =
| deity_of    =
| abode       =
| symbol      =
| consort    =
| parents    =
| siblings  =
| children      =
| mount         =
| other_names   =
| Greek_equivalent =
}}
```

Figure 2.2: Example Infobox of the page related to the god Mercury with the default properties for infoboxes about deities.

2.2.1.3 Categories

Every Wikipedia article should have at least one category. Categories are collections that identify topics in the encyclopedia. It should be noted that although the structure of the Wikipedia categories form a taxonomy, it is not represented by a simple tree of subcategories but, in fact, by a complex graph. This graph allows multiple simultaneous categorizations of topics, which means that one category may have multiple parents. The category "Semantics" is a good example of this complex structure since it is a subcategory

of “Grammar”, “Linguistics”, “Concepts in logic”, “Semiotics”, “Philosophy of language” and others as demonstrated in figure 2.3.

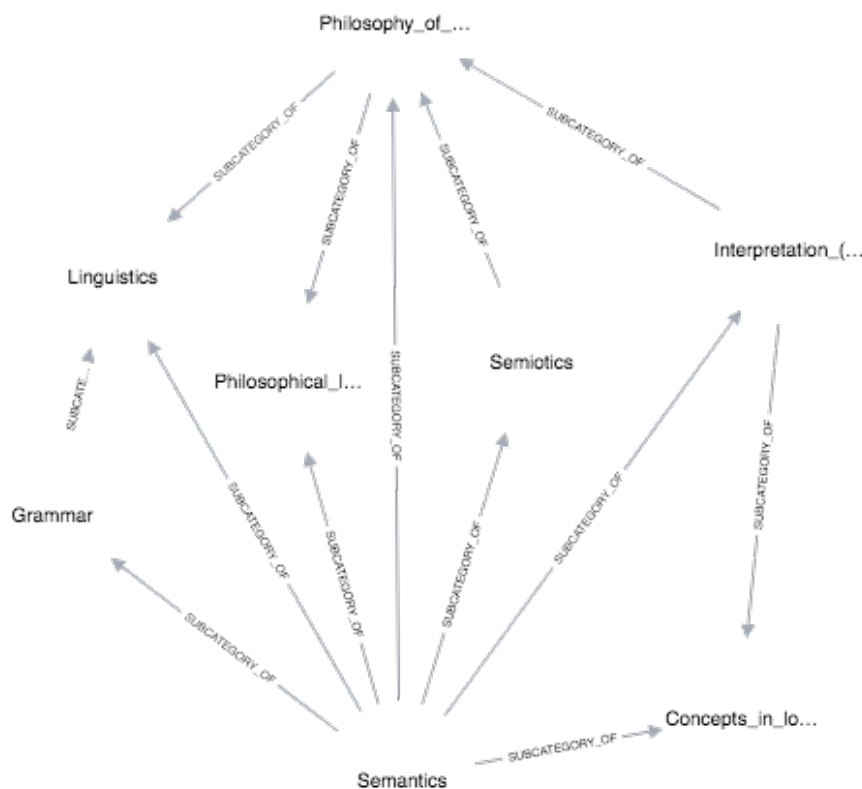


Figure 2.3: Example of an induced graph showing the supercategories of “Semantics” in the Wikipedia Category Graph.

Although not based on semantics, Wikipedia has a set of characteristics, such as the definition of a large number of articles and organization of articles in categories, which make it an essential semantic resource. A simple example, but one that illustrates the complexity of the relationships between Wikipedia categories well, is the “Apple” concept (the fruit), which is directly linked to four categories: “Apples”, “Malus”, “Fruits originating in Asia”, and “Plants described in 1768”. Each of these categories has been added and curated by people who are part of the Wikipedia community. In addition to the explicit knowledge in the directly attributed categories, a vast quantity of implicit knowledge can be inferred by the relations between them, both generically and specifically. Taking the category “Apples” as the starting point, links can be followed to the top of the classifi-

cation system, in what can be perceived as a more generic case: Apples \rightarrow Edible Fruits \rightarrow Edible Plants \rightarrow Food \rightarrow Food and Drink \rightarrow Health. A more specific case can be illustrated by taking the “Malus” category as the origin, and analyzing one of the possible paths to the top: Malus \rightarrow Maleae \rightarrow Prunoideae \rightarrow Rosaceae \rightarrow Rosales \rightarrow Rosids \rightarrow Core Eudicots \rightarrow Eudicots \rightarrow Angiosperms \rightarrow Plants \rightarrow Eukaryota \rightarrow Organisms \rightarrow Life. These are examples capture small fragments in Wikipedia’s categorization structure for the “Apple” concept. The complete structure involves 33 different categories and 42 different relations between them (See figure 2.4).

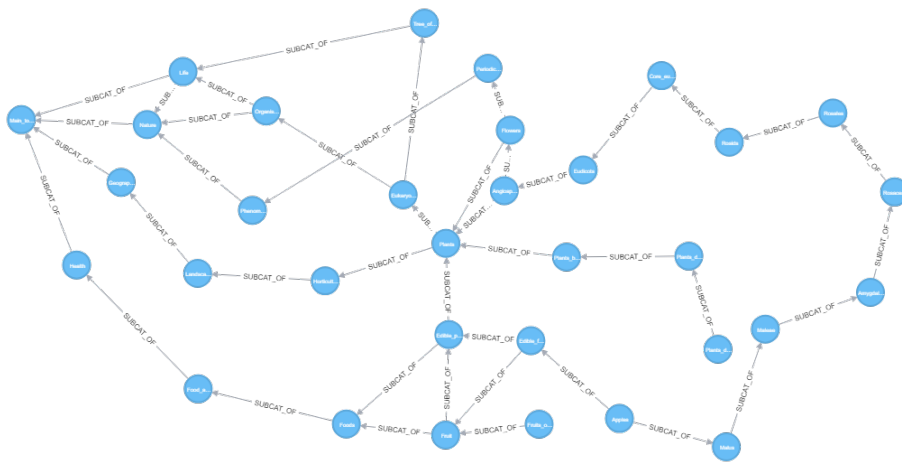


Figure 2.4: Example of an induced graph showing the categories and relationships for the Entity Apple towards Main topics

2.2.2 The Wikipedia Category Graph (WCG)

Regarding the reduction of dimensionality, a proposed method consists of navigating the WCG from each category extracted related to the entities obtained from the text-based resources towards the top of the graph, by all the shortest paths between the category and a set of top-level categories.

The WCG mentioned above is a set of almost 1,500 categories, describing a broad domain of knowledge and ranging from the very precise, such as “Lists of Canadian network television schedules”, to the very general, such as “information”. The categories are connected by hypernym relationships, with a child category having an “subcategory-of” relationship to its parents in the direction of the relationship. However, the graph is not

strictly hierarchic: shortcuts exist in the connections (i.e., starting from one child category and going up two different paths of different lengths to reach the same parent category) as well as loops (i.e., beginning from one child category and going up a path to reach the same child category again).

Given the complexity and dimension of the WCG, a graph-theoretic analysis was carried out and is described in Chapter 3.

3. The Wikipedia Category Graph

Exploiting the underlying structure of Wikipedia requires a way of representing it. In this chapter, a topological analysis of this structure represented as a directed graph has been performed.

The primary goal of the analysis expressed in this chapter is understanding the organization of the Wikipedia body of knowledge regarding its structure of categories, as well as the possibilities and challenges that emerge from decoding this underlying structure.

The inspiration for this chapter comes from the work presented by Zesch and Gurevych [79], where they showed that the WCG is a scale-free, small-world graph, similar to other semantic networks such as WordNet [48] or Thesaurus.com¹. They concluded that the WCG could be used for Natural Language Processing (NLP) tasks, where other semantic networks have been traditionally employed. Although their work has been useful in supporting many types of research in the past years, the analysis was restricted to the German version of Wikipedia as it stood in 2007.

This thesis contains an up-to-date review of a more recent version of Wikipedia, to obtain insights on how the structure of the WCG can influence the proposed method and guide the development of future work.

¹<https://www.thesaurus.com/>

3.1 The Category Graph

A category graph is a way of representing existing relationships between categories, such as which subcategories can be reached from one another. Figure 3.1 illustrates how the categories are organized and connected if they are represented as a graph. The nodes in the graph represent categories, and the edges represent the relationships between them.

The graph illustrated in figure 3.1 is directed with each edge representing the relationship between a pair of categories (e.g., C1-1 is a subcategory of TC-1 since the arrow points from C1 to TC1). TC-1, TC-2, and TC-3 represent the top-level categories of Wikipedia. At the time of writing, there are a total of 19 top-level categories in the English Wikipedia that summarize the entirety of the body of knowledge encoded within its articles (namely Arts, Culture, Games, Geography, Health, History, Humanities, Industry, Law, Life, Mathematics, Matter, Nature, Philosophy, People, Reference Works, Religion, Science and Technology, and Society). The HC1 represents a hidden category. Hidden categories are not displayed in the Wikipedia articles to general users, even if the article is placed under that category. These categories are mainly used for internal organization and do not provide any meaning - for that reason, they have been omitted from the graph.

The details of information extraction from Wikipedia, and the ways in this information was filtered in order to assemble and represent the WCG for carrying out the analysis (as described in this chapter) are reported in Appendix A. The use of the graph as the basis for the proposed approach is also described.

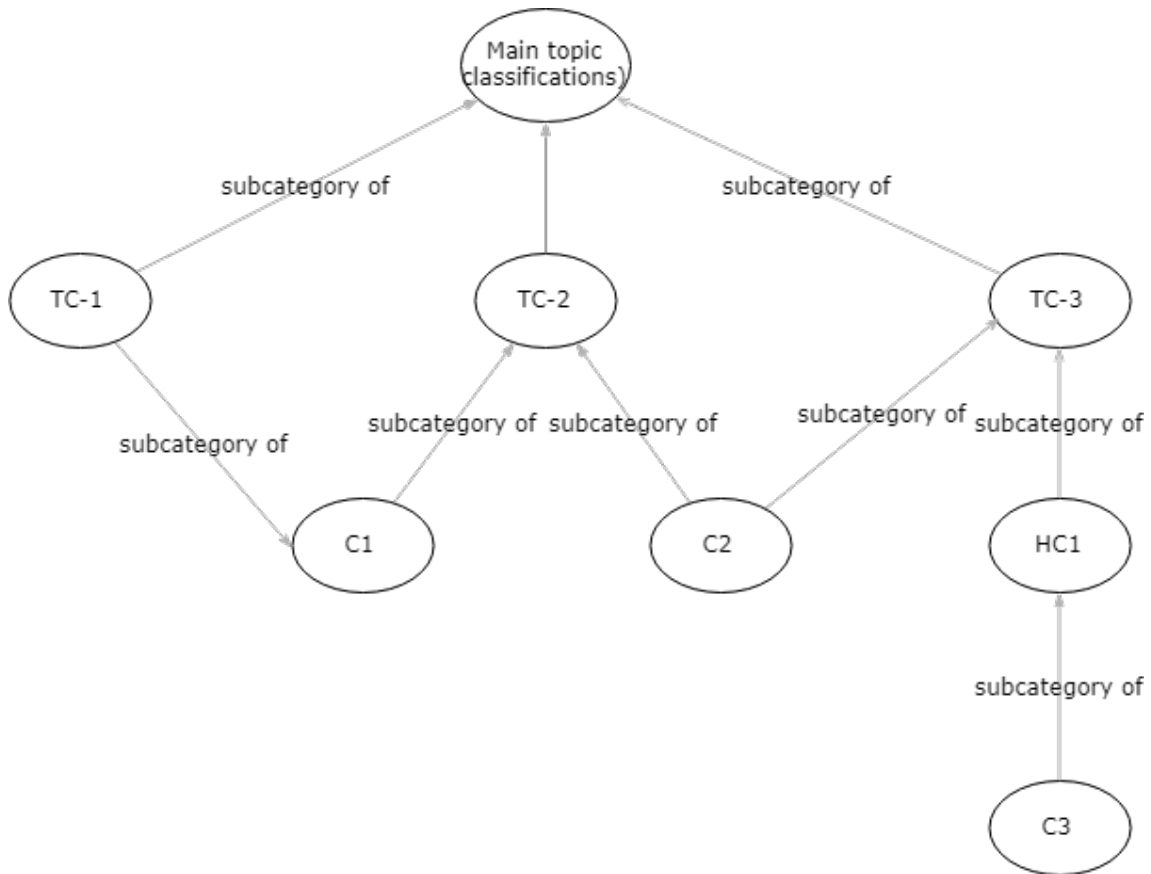


Figure 3.1: Simplified example of the underlying structure of the WCG

3.2 Small-world networks (SW)

According to Steyvers and Tenenbaum [66], interest in studying the small-world phenomenon originated with social network studies when the results suggested that any two people were, on average, separated by only a few acquaintances or friends (so-called “six degrees of separation”). While the finding of very short path lengths between random pairs of nodes in a network may seem unexpected, the phenomenon is well-described by even the simplest models of random graph theory, such as the one by Erdős and Rényi [10]. In an Erdős and Rényi random graph with n nodes, any pair of nodes is connected by an edge with probability p . When p is sufficiently high, the whole network becomes connected: the average path-length, L , grows logarithmically with n , the size of the network.

Watts and Strogatz [75] formally defined small-world networks as a class of networks that are highly clustered, like regular networks, yet have small characteristic path lengths,

like random graphs. These characteristics result in networks with unique properties of regional specialization with efficient information transfer.

Most real-world networks, such as the World Wide Web (WWW), networks of scientific collaborators, and metabolic networks in biology do not have the homogeneous distribution of degree (the degree of a node is the number of neighbors a node has) that regular or random networks have. The number of connections each node has varies considerably in most networks, and they are positioned somewhere between regular and random networks. [66]

Figure 3.5 displays the difference between distribution of nodes in a regular network (3.2, a small-world network (3.3) and a random network (3.4).

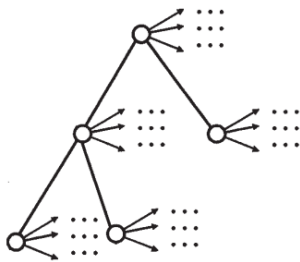


Figure 3.2: A tree-structured hierarchy



Figure 3.3: Scale-free small-world graph

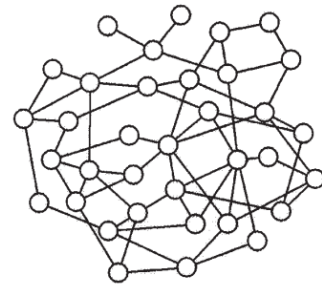


Figure 3.4: An arbitrary, unstructured random graph

Figure 3.5: Structures of semantic networks adapted from Steyvers and Tenenbaum [66]

Steyvers and Tenenbaum [66], analyzed different large-scale semantic networks (such as WordNet [48] and Thesaurus.com) regarding Sparsity, Connectedness, Path-Lengths, Clustering coefficient and degree distribution, concluding that all of them exhibit a scale-free pattern. These metrics are described and analyzed in the context of the WCG below.

3.2.1 Sparsity and Connectedness

The WCG assembled for the experiments contains 1,475,806 nodes and 4,091,417 edges. Each node represents a category and each edge represents a relationship of the type “Subcategory Of”.

The density D of a graph G is the ratio of edges in G to the maximum possible number of edges, defined in a directed graph as

$$D = \frac{|E|}{|V|(|V| - 1)} \quad (3.1)$$

where V is the set of nodes and E is the set of edges.

A graph is said to be dense when the number of existing edges is close to the number of possible edges. In the WCG, the density is $1.878519 * 10^{-6}$. As in the semantic structures analyzed in [66], on average, a node is connected to only a tiny percentage of other nodes.

In the WCG examined for this thesis, the largest connected component contained 99.23% of the total nodes. Despite the sparsity, networks that are not random form one large connected component: from one node, any other node can be reached by some associative path.

3.2.2 Path lengths

A path in a graph is a sequence of alternating nodes and edges that starts with a node and ends with another node in such a way that adjacent nodes and edges in the sequence are incidental to each other [52]. Nodes or edges can appear in the same path multiple times, and the number of edges in a path is the length of that given path. If a graph is connected, then any node can be reached via a finite-length path starting from any other node. The shortest path between a pair of nodes is called a geodesic path, and there can be more than one such path.

The average path length, a concept in the field of network topology, is defined as the average number of steps in the shortest paths for all possible pairs of nodes in the graph. In directed graphs, the average path length is calculated as follows:

$$l_G = \frac{1}{2 * n * (n - 1)} \cdot \sum_{i \neq j} d(v_i, v_j) \quad (3.2)$$

where $d(v_i, v_j)$ denotes the shortest distance between v_i and v_j and n is the number of nodes in the graph G .

If two nodes are disconnected (i.e., no path exists between them), the path length between these nodes is infinite. Consequently, if a graph contains disconnected components, the average shortest path length l_G tends to infinity. Given that the WCG is not completely connected, to avoid infinity, the average shortest path length was calculated for the largest connected component. As a result, the l_G is 20,9343. The shortest path length distribution is displayed in figure 3.6 where the y-axis represents the number of nodes and the x-axis represents the number average path length.

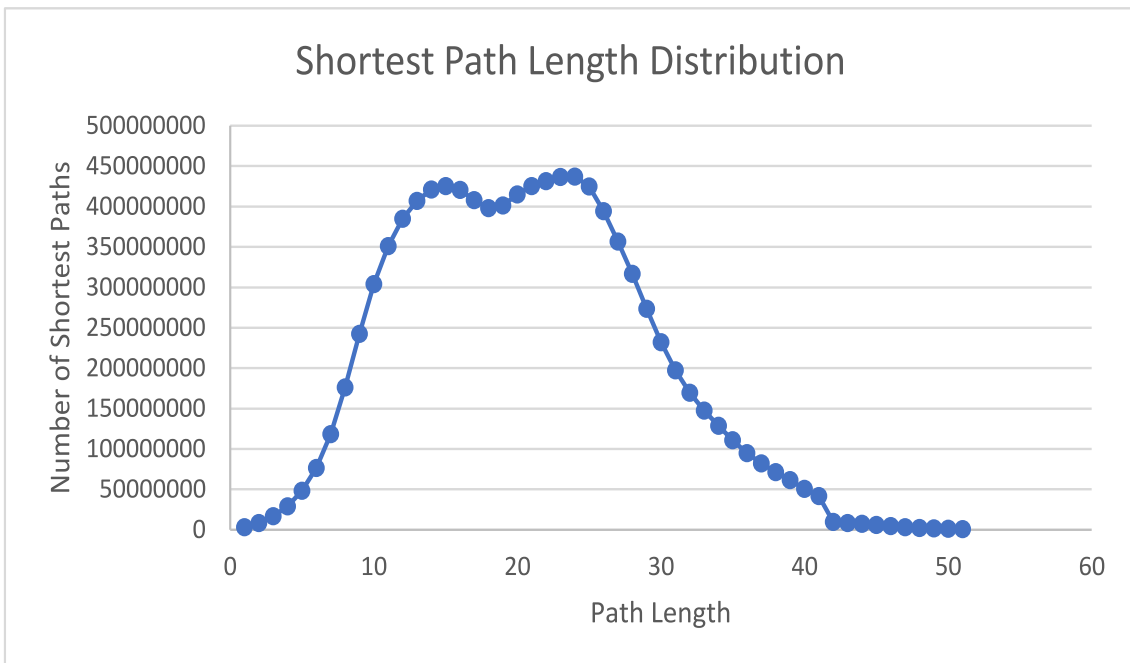


Figure 3.6: Shortest path length distribution on the WCG

3.2.3 Clustering Coefficient

Another important metric for understanding the topology of a graph is its clustering coefficient. The clustering coefficient of a node represents the probability that if two of its neighbors are randomly chosen, they will also be connected by an edge. More precisely, if a node has t neighbors, then there are $t(t-1)/2$ possible edges that connect those neighbors. The local clustering coefficient for a node is then given by the proportion of edges between nodes within its neighborhood divided by the number of links that could

exist between them.

The clustering coefficient C of the whole graph G is the average of the local clustering coefficients $C(v)$ of all nodes $v \in V$.

Figure 3.7 illustrates how the clustering coefficient of a node is calculated.

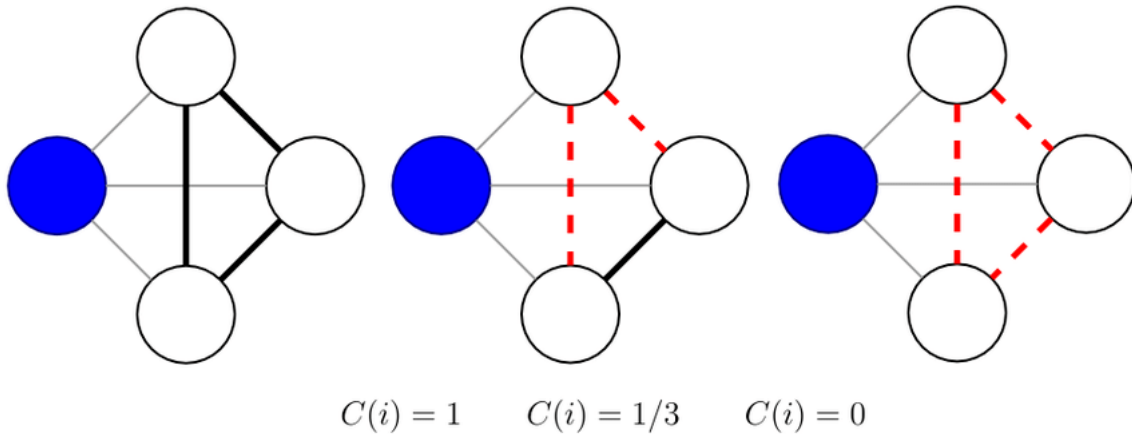


Figure 3.7: Example of how the clustering coefficient of a node is calculated based on the probability of the neighbors of the node i are also neighbors among themselves

The WCG has a clustering coefficient of 0.0461. This value of clustering is orders of magnitude larger than can be expected from random graphs of equivalent size and density. The same phenomenon was observed in the networks analyzed in [66]. The impact of clustering in scale-free networks is described below.

3.2.4 Degree Distribution

The indegree of node v is the total number of connections into node v ; the outdegree of node v is the total number of connections coming out from the node. The indegree and outdegree of the graph is the average of the degree of each node presented in the graph.

Large graphs such as the WCG are complex structures, as the connections among the nodes can present complicated patterns. While studying complex networks, it is common to develop simplified measures that capture some elements of the structure. The degree distribution of a complex network is often described in this context.

The degree distribution of a graph is the probability distribution that a randomly chosen node will have a degree k . In directed graphs the degree distribution is a two-dimensional distribution, so that $P_{\text{deg}}(k^{\text{in}}, k^{\text{out}}) =$ the portion of nodes in the graph with indegree k^{in} and outdegree k^{out} .

There is a large class of so-called scale-free networks, characterized by a highly heterogeneous degree distribution, which follows a power-law². They are called scale-free because zooming in on any part of the distribution does not change the shape of the network: there is a few, but a significant number of nodes with many connections and there is a trailing tail of nodes with a very few links at each level of magnification [66].

A characteristic of a scale-free network, derived from its degree distribution, is its tolerance to mistakes. If a random node is disconnected from the network, the highest probability is that this node has a low degree, causing a little impact for the interconnectivity of the remaining network.

Small-world networks tend to show a power-law distribution which means that the fraction $P(k)$ of nodes in the graph having k connections to other nodes varies as a power of some attribute α , as in:

$$P(k) = Ck^{-\alpha} \quad (3.3)$$

To define the best with the best fitting power-law curves, the method applied is defined in [9] as

$$\alpha = 1 + n \left[\sum_{i=1}^n \ln \frac{x_i}{x_{\min}} \right]^{-1} \quad (3.4)$$

where x_{\min} is a lower cutoff, below which the power-law cannot be observed. The parameters observed are shown in table 3.1.

| | α | x_{\min} |
|-----------|----------|------------|
| indegree | 2,4124 | 10 |
| outdegree | 4,5603 | 12 |

Table 3.1: Power-law parameters of the WCG

²A relationship between two variables such that one is proportional to a fixed power of the other.

Figure 3.8 shows the degree distribution for the WCG in $\log\text{-}\log^3$ plot. The x-axis shows the degree while the y-axis shows the count of nodes with such degree.

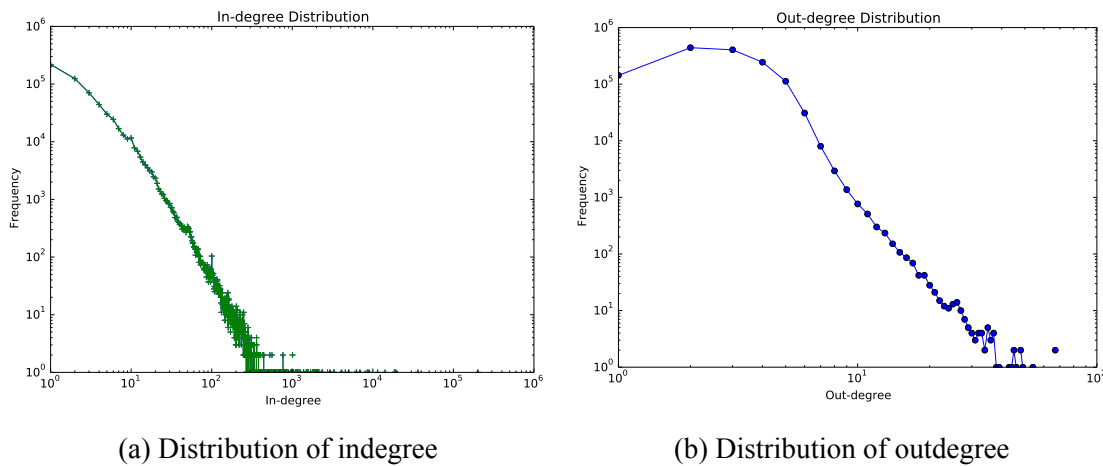


Figure 3.8: Indegree and outdegree distributions

Based on the analysis of the WCG degree distribution, as presented by the plot in figure 3.8 and the exponent displayed in table 3.1, it is demonstrated that the WCG nodes follow a power-law distribution. Hence it can be considered as a scale-free network. The same was reported in [66] regarding other large semantic networks.

3.2.5 Empirical demonstration of small-worldness

Based on the metrics evaluated by Steyvers and Tenenbaum [66] and reported in this chapter, it is possible to infer that the structure of the WCG is very similar to the small-world networks. However, to support this statement, it is demonstrated based on a mathematical method below.

Humphries and Gurney [28] described a mathematical method to determine whether or not a graph can be considered a small-world based on the comparison of the graph parameters (WCG) with a random graph with the same proportions.

Formally, a graph G with n nodes and m edges is said to be a small-world if it has a similar path length but greater clustering of nodes than an equivalent Erdős-Rényi (E-

³ A two-dimensional graph of numerical data that uses logarithmic scales on both the horizontal and vertical axes.

R) random graph with the same m and n . Let L_g be the mean shortest path length of G and C_g its clustering coefficient. Let L_{rand} and C_{rand} be the corresponding quantities for the corresponding E–R random graph. According to the empirical experiments performed in [28], a graph G is said to be a small-world network if $L_g \geq L_{\text{rand}}$ and $C_g \gg C_{\text{rand}}$. Table 3.2 shows the summarized values for the WCG at the centre of this thesis, and for a random graph with the same number of nodes and edges.

Table 3.2: Empirical demonstration of small-worldness of the WCG.

| | L_g | L_{rand} | C_g | C_{rand} |
|-----|---------|-------------------|--------|-------------------|
| WCG | 20.9343 | 19.5389 | 0.0461 | 0.00000003 |

The WCG exhibits small-world behavior, with an average shortest path length close to that of a random network of the same size. The clustering coefficient, however, is orders of magnitude higher than in the random graph.

3.3 Final Consideration

In this chapter, the representation of the categories of Wikipedia and the relationships between them in the form of a directed graph has been described. An analysis of the topology of the graph was performed, empirically demonstrating that, as with other large-scale semantic networks, the WCG can also be characterized as a small-world and scale-free network.

Challenges encountered in this analysis included the considerable computational power required by the processes of both the information extraction and the analysis. This is likely to be the main reason why most of the existing literature on the topic has tended to focus on the analysis of semantic networks that are much smaller, but also quickly out of date.

This analysis supports some critical decisions related to the proposed method. The aggregation of categories consists of navigating the Category Graph from each category extracted from the named entities in the text towards the top-level categories. Considering that the WCG is a small world, navigating towards all paths would be nonsensical, since

each category can be reached from any other. However, because the WCG resembles other semantic networks commonly used in NLP and IR applications, the shortest paths can be used, as in this type of graph they carry a robust semantic relation [47].

As a scale-free network, the structure of the the WCG is notably fault tolerant. The categorization process in Wikipedia is a continuous work-in-progress, as users edit, remove, and add categories frequently. However, the probability that an edit spoils the overall structure of the graph is very low since most categories do not have a high degree of connection to the whole graph.

In addition to the fact that it is scale-free, WCG has another significant advantage for categorization: the subjects are divided into well-connected neighborhoods, and the neighborhoods are interconnected to some extent. In practice, this means that specific knowledge about a subject is presented in a well-connected way, while transversal knowledge can also be captured from this structure.

4. Fundamental Concepts

This chapter introduces the theoretical concepts used to contextualize the applied approach for classifying documents on the Web. The concepts related to Information Retrieval (IR) and Documents Representation are essential to understanding the tasks applied in the proposed solution to extract the entities and the categories from text-resources online. An overview of the concept and methods for the automatic classification of textual resources, the primary goal of the research presented in this thesis, is presented below.

4.1 Information Retrieval

IR systems are mostly known for their searching ability, where a user states an information need and the system provides the user with a response to this information need in return. IR is a large academic field and encompasses several topics such as browsing or filtering documents, processing of retrieved documents and clustering or classifying documents according to their content. In [39], Manning defines IR as finding material (usually documents) of an unstructured or semi-structured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Structured resources are machine-readable resources that encode relationships of various types according to the level of information [25]. They are of high quality as they are built from the knowledge of domain experts, lexicographers, and linguists, but limited because they require significant efforts in creation and updating. Because they are built

manually, they depend on the availability of experts to extend their coverage and to keep them up to date on recent events. Moreover, knowledge encoded in one language is not transferable to others, requiring a new effort for each new language.

The most common structured resources are:

- Thesaurus: collections of related terms;
- Taxonomies: Hierarchical structures of classification of terms;
- Ontologies: knowledge models that include concepts, relations of different types, rules and axioms.

Unstructured resources are collections of texts that have no formalized knowledge and are machine readable only as sequences of characters and words. Different statistical models can extract knowledge from unstructured collections, and the vast number of texts available on the WWW enables the construction of knowledge bases with extensive coverage. However, they are limited by the lack of texts that demonstrate common-sense knowledge [25]. Also, statistical models are not able to issue knowledge with quality equivalent to the resources built by specialists.

The limitations of unstructured resources are complementary to those of structured resources. While unstructured resources enable broad coverage with low quality, structured resources have high quality but low coverage. Semi-structured resources constructed collaboratively on the WWW encode the knowledge voluntarily made available by users of these resources, covering different areas of expertise and having quality comparable to that obtained from specialists. Examples of semi-structured resources are Wiktionary, Flickr, Twitter, Yahoo! Answers and, Wikipedia [25].

Nowadays, most of government, industry, business, and other institutions information are stored electronically on the Web as semi-structured or e-structured resources.

In this context, the existence of IR systems becomes indispensable in assisting users

in the process of locating relevant information in collections of unstructured or semi-structured data (e.g., web pages, documents, images, videos).

A system can increase its precision when addressing the users' information need if it indexes well-represented features of the text. As the collection of documents expands, automatic techniques that can extract these features become crucial. Text classification techniques can provide an representative output of a research document, allowing for IR systems to handle the indexing and retrieval process.

In order to reduce the complexity of the documents and make them easier to handle for IR systems, each document has to be transformed from the full-text version to a compact representation, which describes its contents [42]. This task of document representation is crucial for text classification approaches.

4.1.1 Documents Representation

Plain texts are usually not used directly by classification algorithms. The documents are processed and transformed to represent the semantic content of the text, optimized for numeric processing.

The Vector Space Model (VSM) is a simple, traditional and practical model that makes it possible to represent documents as vectors and to perform any algebraic operations to compare them [58].

In this method, the documents of a collection D are represented in the VSM as points in a multidimensional Euclidean space, where each dimension corresponds to a distinct term in that collection. The set T of distinct terms of collection D , called vocabulary collection of D , is obtained in a process called lexical analysis.

This type of representation is widely used in IR, in tasks of textual retrieval, ordering of documents by relevance (ranking), and text classification tasks. The use of vector representation makes the use of any algebraic operation applicable to this type of structure possible, enabling comparisons between two documents, as explained by Salton [59].

Each term of the set T can be composed of only one word (unigrams), several words (bigrams, trigrams or n -grams) or sentences, and has an associated weight to determine its degree of importance [58].

Given a document $d_i \in D$, this document is formally represented in the VSM as follows $d = w_{i1}, w_{i2}, w_{i3}, \dots, w_{i|T|}$, where T is the vocabulary set of the collection D and w_{ij} ($1 \leq j \leq |T|$) is the weight of the term t_j in the document d_i , such that $w_{ij} = 0$ if the term t_j does not occur in the document d_i .

To represent the text in the VSM, most text classification methods use the Bag of Words (BOW) approach to represent documents [34] [39]. The categorization takes into account the presence or the absence of key terms in the document-terms matrix [61].

The reasons for using this approach is the simplicity, efficiency and relative effectiveness of the BOW paradigm. However, the BOW method fails to take into account relevant aspects of the text that is being represented. Semantic relationships between key terms are ignored, as well as the order in which the terms appear [14, 34]. The BOW approach ignores essential semantic relations between the terms [26].

Elements of bipartite words such as “White House” or “Bill Gates” are represented in the BOW as unrelated words. In analyzing the BOW representation of a given document in which the words “bill” and “gates” occur, one might suggest that the document talks about accounting for a construction firm (the word “bill” for accounting, and construction from “gates”). For a computer program, it would be challenging to associate these words. Nevertheless, if the representation of the same document contains the set of words “Bill Gates” as a term, it would be easier for the classifier to make a correct association. [4].

As a consequence, if two documents use different sets of keywords to describe the same topic, they can be classified as being of different categories, even if the keywords used by both are synonymous or semantically associated in some other form [27].

Among the alternative representations that use features of the text itself, most common are those that use sequential co-occurrence of n terms (n -grams) and non-sequential

co-occurrences of n terms (term sets). Other approaches have explored features that are not directly extracted from the text. The growing interest in Features Generation (FG) techniques known as Document Expansion or Document Enrichment, through which new terms are added to documents, enhancing the BOW representation by inserting more information in the document-term matrix [14] is an example of this. Numerous methods that use FG have achieved verifiable results in text classification through the extraction of semantic relations such as synonymy, hyponymy and associative relations between concepts, present in encyclopedias, thesauri, ontologies, web pages and other sources [14, 15, 26, 73, 74].

Since many texts used in the Web (search queries, tweets, questions, and answers) are short, unstructured and ambiguous, there is a need for a methodology capable of analyzing short text semantically, detecting possible entities present in the sentence, disambiguating between terms, and overcoming the gaps of tradition methods. The approach described in this thesis employs the named entities found in the text as the basis for the representation, enabling the capture of semantic information related to the documents present on the Web.

4.1.2 Named Entity Recognition

Grishman and Sundheim [22] defined the task of NER, as one “which involves identifying the names of all people, organizations and geographic locations in a text”, whilst also involving the identification of date expressions, time, monetary values and percentages.

The NER task has been researched under several names over the years, for example Wikification [56], Grounding [35] or Named-entity disambiguation [24]. The common approach can be generalized into the following steps: finding named entity mentions in a given text; generate a set of candidates for each mention; select the best candidate; and link the selected mention to the corresponding entry in the knowledge base.

In its most common form, the NER task recognizes a predefined number of semantic categories, such as those defined in [22]. However, it has also been successfully applied to specific domains such as biology [7] and geology [65], where a more substantial number

of domain-related categories are used. Another prevalent form is the use of linguistic rules for the recognition of entities. In this approach, the rules are manually coded from grammatical and domain knowledge, requiring specialization in both to obtain good results [51]. The use of linguistic rules restricts its applicability to documents written in the language for which the rules were codified, making it unusable with other languages.

The NER task consists of two phases: i) the annotation of the grammatical classes of the text, and ii) the annotation of the names with the semantic category.

The annotations identify the grammatical class for each word in the text. If the word is an entity name, they identify its type. The quality of the algorithm is dependent on the annotator's ability to identify the names correctly, and it is limited to the types of entities used in the corpus.

Figure 4.1 shows an example of the NER task in a post extracted from a social network¹. The word “Michelle” was identified as the Person Michelle Obama and the Words “New York” were identified as the Place New York City.

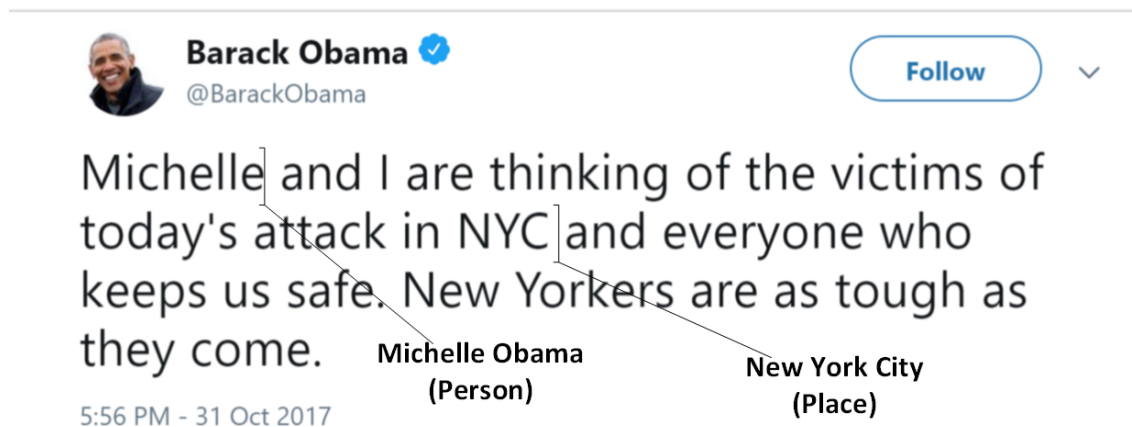


Figure 4.1: Example of the NER task on a text extracted from a social network.

4.2 Automatic Text Classification

Automatic text categorization is the activity of assigning text in natural language with categories from a predefined set according to their content. [77].

¹<https://twitter.com/BarackObama/status/925526988659548160>

This task can be formalized as follows: let $D = \{d_1, d_2, d_3, \dots, d_n\}$ be a finite set of documents and $C = \{c_1, c_2, c_3, \dots, c_n\}$ a finite set of predefined categories.

The problem is finding a function $f : D \times C \mapsto \mathbb{R}$ that assigns a score s for each pair $\{d_i, c_i\} \in D \times C$, where the membership value of s specifies the degree of relevance of the category c_i to the document d_j .

Two different types of text categorization task can be identified depending on the number of categories that could be assigned to each document. The first type, in which precisely one category is assigned to each $d_j \in D$, is named as the single-class (or non-overlapping categories) text categorization task. The second type, in which any number of categories from zero to $|C|$ may be assigned to each $d_j \in D$, is called the multi-class (or overlapping categories) task.

According to Sebastiani[61], the definition of a text classifier can be summarized in the following steps:

1. Acquisition of documents belonging to the domain (i.e., a collection of documents in D);
2. Creation of a vocabulary of T terms $\{w_1, w_2, w_3, \dots, w_T\}$ that will be used in the representation of documents. This step involves preprocessing, such as lexical analysis (removal of digits, punctuation marks), removal of stopwords (articles, prepositions), and the stemming of words (reduction of the word to its radical) among other text operations;
3. Creation of the initial representation of documents with the definition of the set of attributes that describe them. Each attribute may be merely a Boolean value that indicates whether or not a given vocabulary term exists in the document (i.e., Boolean representation). Each attribute can also correspond to a numerical weight associated with a given term, indicating its relevance to the document being described.
4. Dimensionality Reduction, where the M most relevant attributes of the initial rep-

resentation are selected (with $M < T$). This step can be done using different criteria for the selection of attributes, such as Information Gain, Mutual Information, $\tilde{\chi}^2$ statistic, and others [78].

5. Induction of the classifier from a training set.
6. Performance evaluation from a test set. Metrics used to evaluate classifiers include precision, accuracy, recall, and F-measure. Further details regarding these evaluation metrics can be found in the description of the experiments (chapter 6).

The methods that are used in text classification are frequently the same as those used in the more general area of IR, where the goal is to find documents or sections within documents that are related to a particular query. Text classification methods are essential to finding relevant information in many different tasks that deal with large volumes of text-based information, such as finding Internet pages on a given subject, finding answers to similar questions that have been answered before, or classifying news by subject or newsgroup, among others. In each case, the goal is to assign the appropriate category or label to each document that needs to be classified.

Over the last few years, a vast number of algorithms have been proposed for text classification using machine learning. Among them, one can cite the naive Bayes [40], K-nearest Neighbors (KNN) [57], Support Vector Machines (SVM) [29] and rule learning algorithms [64]. Research focusing on document classification usually reports on performance comparisons between different available algorithms.

The approach proposed in this thesis is a multi-label classification method, capable of assigning different degrees of membership for a document regarding each one of the categories available. Multi-label classification problems can usually be reduced to a particular case of single-label classification. To make the results suitable for comparison with other approaches and with the judgment of humans, multi-label classifications were converted to a single-label assignment by considering only the categories with the highest degrees of membership.

5. Methods

This chapter presents in detail the steps of the approach applied to the extraction and representation of document features based on the WCG.

5.1 Approach

The rich structure of the WCG has contributed to making it a large and meaningful semantic taxonomy. The main goal of the research reported on in this thesis is to take advantage of this body of knowledge by automatically categorizing text-based content on the Web following the collective knowledge of Wikipedia contributors. A processing chain to generate a generic categorization was developed based on three steps:

1. Text annotation;
2. Categories extraction; and
3. Document Representation.

The relationships between Wikipedia Categories have been considered as a directed graph. Let $G=(V, E)$ be a graph, where V is the set of nodes representing Wikipedia categories, and E is the set of edges representing the relationships between them.

For ease of comprehension, the method will be illustrated by way of a running example in a possible real application scenario.

Imagine that you are the editor of a news site and have received a complex article from the journalist to be published. Your task is to define which categories to attribute to the article so that the content is well represented and allows users to find it on the site quickly.

As an example for this scenario, let us use an excerpt from the article entitled “Market Totalitarianism in North Korea”, taken from The New York Times¹ of May 3, 2017:

*“(...) Post-Communist, postindustrial, kleptocratic dynastic regime of North Korea may become the crown jewel of the new axis-of-tyranny ideology (...)”*²

5.1.1 Text Annotation

Documents on the Web are primarily unstructured data, which hinders data manipulation and the identification of atomic elements in the texts. To alleviate this problem, Information Extraction (IE) methods, such as NER are employed. These methods automatically extract structured information from unstructured data and make it possible to link them to external knowledge bases.

DBpedia describes itself as a “crowd-sourced community effort to extract structured content from the information created in various Wikimedia projects”³. This structured information resembles an Open Knowledge Graph (OKG) which is available for everyone on the Web. A knowledge graph is a particular kind of database which stores knowledge in a machine-readable form and provides a means for information to be collected, organized, shared, searched and utilized.

In the context of this thesis, DBpedia was chosen as the knowledge base because it covers many domains (Science, Arts, Politics, History, Geography, Health and Nature, among others). Another reason is its constant evolution. Since the knowledge in DBpedia is extracted from Wikipedia, the Knowledge Base is also continuously updated by the contributors. DBpedia is also available in different languages and can be accessed either

¹<https://www.nytimes.com/>

²<https://nyti.ms/2py3z4r>

³[urlhttps://wiki.dbpedia.org/about](https://wiki.dbpedia.org/about)

by an endpoint⁴ or being installed in a local machine, making it faster to process a vast amount of data.

Based on a comparison made by Gangemi [19], the decision to use the DBpedia Spotlight tool⁵ for entity extraction and linking to DBpedia was made. Although there are some options (such as AIDA⁶ or Alchemy⁷) that outperform Spotlight for the task of NER, they fail to meet other criteria. For instance, AIDA is directly linked to YAGO and Alchemy is a paid API with limited access.

DBpedia Spotlight is a system for automatically annotating text documents with DBpedia URIs. It contains Wikipedia's encyclopedic knowledge of some 3.5 million resources, where nearly half of the knowledge base is classified according to the following ontologies: people, organizations, and places [45]. DBpedia Spotlight was used to extract and enrich entities found in Web resources.

To return to the running example:

*"(...) Post-Communist, postindustrial, kleptocratic dynastic regime of North Korea may become the crown jewel of the new axis-of-tyranny ideology (...)"*⁸

After using DBpedia Spotlight to extract the concept from the excerpt, the entities linked to DBpedia as shown in table 5.1 were obtained.

⁴<http://dbpedia.org/sparql/>

⁵<http://dbpedia-spotlight.github.io/demo/>

⁶<https://github.com/codepie/aida>

⁷<https://www.ibm.com/watson/alchemy-api.html>

⁸<https://nyti.ms/2py3z4r>

Table 5.1: Entities extracted from the text and their respective links to DBpedia concepts.

| Entity in Text | Link to Dbpedia |
|----------------|---|
| postindustrial | http://dbpedia.org/resource/Post-industrial_society |
| kleptocratic | http://dbpedia.org/resource/Kleptocracy |
| dynastic | http://dbpedia.org/resource/Dynasty |
| North Korea | http://dbpedia.org/resource/North_Korea |
| tyranny | http://dbpedia.org/resource/Tyrant |

5.1.2 Categories Extraction

Taking the entities found in the previous step as a starting point, the categories extraction step begins by traversing the entity relationships to find a more general representation of the entity, i.e., their categories. All categories associated with the entities identified in the source of information are extracted.

For instance, for each extracted and enriched entity in a Web resource, the proposed methodology explores the relationships through the predicate [dcterms:subject], which by definition represents the categories of an entity. To retrieve these topics, the SPARQL query language was used for querying Resource Description Framework (RDF) over the DBpedia SPARQL endpoint, navigating up in the DBpedia hierarchy to retrieve broader semantic relations between the entities and their topics.⁹

Using a SPARQL Protocol and RDF Query Language (SPARQL) query that retrieves all categories [dc:subject predicate] associated with the entities listed in table 5.1 the categories listed in table 5.2 were obtained.

⁹Note that an entity/concept can be found in different levels of the hierarchical categories of DBpedia. Hence this approach would lead us to retrieve topics in different category levels.

Table 5.2: List of all entities extracted from the example test and the categories associated to them

| Entity | Categories |
|-------------------------|--|
| Post-industrial society | Postindustrial society Information economics Postmodernism Social philosophy Technology in society Theories of history |
| Kleptocracy | Forms of government Political corruption Political terminology |
| Dynasty | Royal families History-related lists Monarchy |
| North Korea | 1948 establishments in North Korea Communist states Countries in Asia East Asian countries Korea Korean-speaking countries and territories Member states of the United Nations Military dictatorships North Korea Northeast Asian countries One-party states Republics Socialist states States and territories established in 1948 Totalitarian states |
| tyranny | Ancient Greek government Ancient Greek titles Ancient Roman government Positions of authority Ancient Greek tyrants |

5.1.3 Representation of Document

The goal of this step is to determine how the resource page being tagged is related to a more generic subset of Wikipedia categories.

In the top of Wikipedia categorization structure, under the “Contents” category there is the category “Main topic classifications”¹⁰ that has 19 subcategories representing different fields of study. The subcategories of “Main topic classifications” were used as a subset of the Wikipedia Categories in the context of this thesis.

The algorithm 5.1 begins with the categories assigned to entities recognized in the text and generates a categorization based on the frequency of assignments with the top-level categories in Wikipedia, the so-called Main topic classifications, based on the definition in section 4.2.

The approach consists of navigating the Category Graph from each category extracted in the previous step towards the top of the graph by all the shortest paths between the category and the main topics. Based on the influence of each main topic category in the resource being classified, a representation of the document based on the calculated categorization as a multidimensional vector using the VSM is generated.

As a formal definition of this step, let us denote I as the set of categories related to a web resource d , found in the category extraction step. C is the set of all Categories in Wikipedia, and M is the set of categories that represent the main topics. $G = (V, E)$, where $I \subset V$; $C \subset V$; $M \subset V$; and $M \subset C$. The parameter l is defined to indicate the broadest l levels to be considered in the set of M . If l is 1, only the main topics previously defined are considered; if l is 2, any category 1 edge away in the graph is also considered as a main topic. Note that a path is a sequence of graph nodes visited from a given category $c \in C$ to a main topic $m \in M$.

The Category Graph is not a perfect hierarchical structure. It is noisy, contains cycles, and many of the paths from a category to the main topic classifications do not represent

¹⁰https://en.wikipedia.org/wiki/Category:Main_topic_classifications

Algorithm 5.1 Vector Generation

```

1: procedure GenerateVector( $G, M, I, t, w$ )
2:    $E \leftarrow$  a map from a list of categories  $m \in M$ 
3:   for  $i \in I$  do
4:      $S \leftarrow$  the set of shortest paths between  $i$  and any category in  $M$ 
5:     for  $s \in S$  do
6:        $B \leftarrow$  the set of last  $t$  nodes in path  $s$ 
7:       for  $b \in B$  do
8:          $E[b] \leftarrow E[b]$ 
9:       end for
10:    end for
11:  end for
12:  return  $E$ 
13: end procedure

```

a meaningful “is subcategory of” relation. One of the main reasons for this is that users who add categories to Wikipedia pages often assign them to small grained categories in the graph. Most of the time, these editors do not fully understand the internal structure of the graph and fail to follow the guidelines when choosing particular categories. The category graph, as well as the content of the articles, can be changed over time, also changing from the original intent of the author and the meaning of the category assignments.

Since the Category Graph is being used as a taxonomy, the shortest path was used to alleviate this problem. This decision is based on the results of the analysis of the topology of WCG, as described in chapter 3. Each time the source category reaches one of the top-level categories by the shortest paths, the influence of this top category in the composition of the resource classification is updated.

In all experiments described in this thesis, only the subcategories of “Main top classifications” were considered in the set of M (i.e. Arts, Culture, Games, Geography, Health, History, Humanities, Industry, Law, Life, Mathematics, Matter, Nature, Philosophy, People, Reference works, Religion, Science and technology and Society) - the parameter of t was fixed with a value of 1. This results in a 19-sized vector representing the content of a given text-based resource.

Returning to our running example, the induced graph contains 143 distinct categories and 256 relationships and can be seen in figure 5.1.

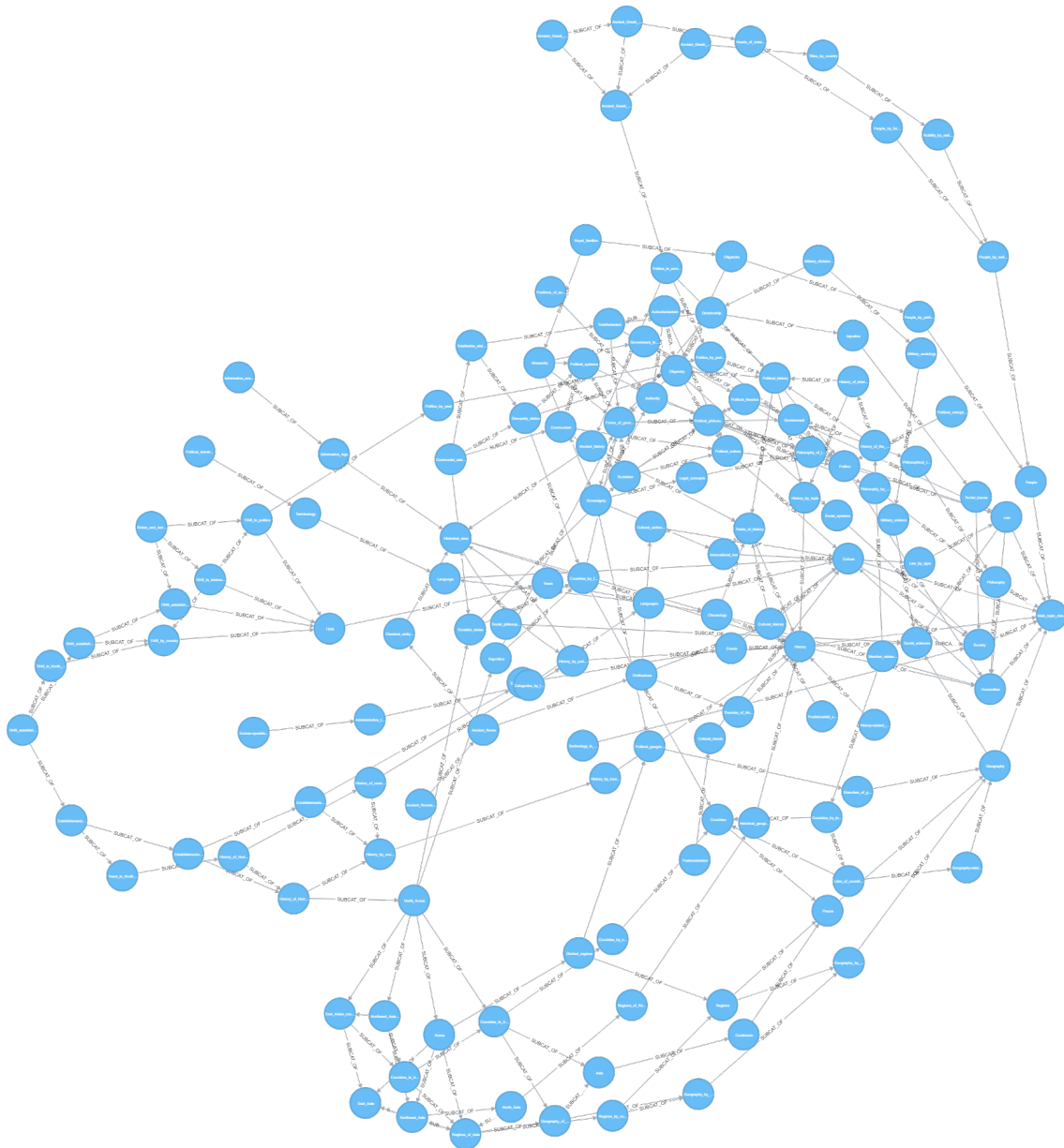


Figure 5.1: An induced graph containing all shortest paths from the categories found and described in table 5.2 and “Main topic classifications” (on the right of the graph)

The vector generated for this example is shown in figure 5.2. Figure 5.2a displays the absolute number of shortest paths used as features for the vector, while figure 5.2b shows the same information normalized. Table 5.3 presents a random example of a path for each top-level category that contributed to the document representation.

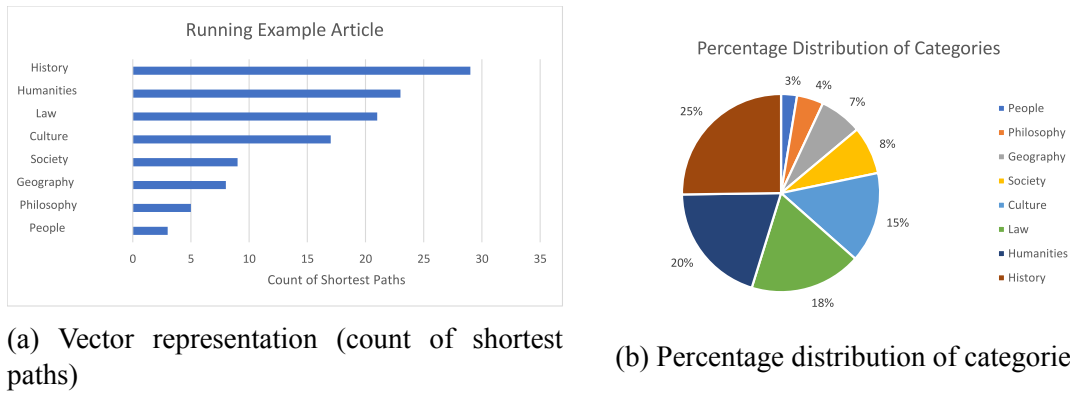


Figure 5.2: Final classification of the running example according to our method

Table 5.3: Example of a path for each top-level category that contributed to the document representation

| Main Category | Shortest path example |
|---------------|---|
| History | Theories of history → History |
| Humanities | Political corruption → Politics → Humanities |
| Law | Political corruption → Politics → Law |
| Culture | Totalitarian states → Totalitarianism → Authoritarianism → Political culture → Culture |
| Society | Military dictatorships → Dictatorship → Oligarchy → Social systems → Society |
| Geography | Korea → Divided regions → Political geography → Branches of geography → Geography |
| Philosophy | Forms of government → Political philosophy → Philosophy by topic → Philosophy |
| People | Royal families → Oligarchs → People by political orientation → People |

For the article used in the running example, it can be inferred that it is strongly related to History, Humanities, Law, and Culture, and also has some weaker relatedness to Society, Geography, Philosophy, and People.

5.2 Final Consideration

This chapter presented the approach applied to the extraction of features and categorization of documents based on the WCG. The method is based on three steps: extracting named entities from text, extracting categories associated with named entities, and finally representing and classifying the document. The prime objective of this methodology is to

generate a classification that can be used by humans, but that can also be applied in automated computational methods. For this reason, the result of classification was presented in two different formats. The vector of feature can be used in machine learning models and can help with automated such as search, retrieval, recommendation, and clustering of information. ON the other hand, the percentage distribution of categories is more tangible from a human point of view. It is important to note that it is difficult to classify content in the Web in an arbitrary way since both in the documents and the Wikipedia structure there is fuzziness, ambiguity, inconsistency and lack of agreement of the contributors regarding some topics. For instance, at the moment of writing, there is no job role for Johann Sebastian Bach (the composer), the contributors cannot agree on what he should be known for. For this reason, the applied methodology does not define a single category for each textual resource, but a percentage distribution of each category concerning how much it contributes to the composition of the whole.

6. Experiments, Results, and Discussion

This chapter presents the methods, results, and discussion of the experiments that were carried out to verify the validity of the proposed approach. It begins with a description of a proof of concept, made by running the classification based on the top-level categories of Wikipedia in posts from ten online Question and Answer (Q&A) communities. An experiment was also run with real users on a crowd-sourcing platform to verify whether the classification generated by the proposed approach was corroborated by humans, the actual users of IR tools.

6.1 Proof of Concept - Q&A Communities

The first evaluation of the approach is a proof of concept aiming to analyze the classification based on the designed method in posts from Q&A communities.

Q&A communities have emerged in the past few years as Web 2.0 has become increasingly popular. They provide a place for users to exchange and share their knowledge explicitly by asking specific questions and by both providing and receiving direct answers within a set of predefined topics and categories.

The volume of questions answered on Q&A sites so far exceeds the number of questions answered by library reference services [62]. Their archives constitute complex and heterogeneous knowledge repositories, presenting a challenge to the organization and retrieval of relevant documents [2].

Stack Exchange¹ is a network of 133 Q&A communities on topics in varied fields. Each community covers a specific theme, where questions, answers, and users are subject to a reputation award process. The decision to use Stack Exchange in the context of the research reported on in this thesis was based on the wide variety of topics covered, and also because the data has been made publicly available in a structured form.

6.1.1 Resources and Methods

An anonymized dump of all user-contributed content on the Stack Exchange network was extracted on August 31st 2017². Each site is formatted as a separate archive consisting of XML files from Posts, Users, Votes, Comments, PostHistory and PostLinks. The Posts files were used as the basis for this experiment. As per the description of the dataset, the property `postTypeId` denotes if the given row in the file is a question or an answer.

Ten representative communities on Stack Exchange were selected to perform this evaluation: Astronomy³, Biology⁴, Chemistry⁵, Christianity⁶, History⁷, Law⁸, Math⁹, Music¹⁰, Philosophy¹¹ and Sports¹². For each row in the `Post.xml` file of each one of these communities, the three steps of the chain described in Section 5.1 were executed. The entities present in each post were first extracted, then linked to their categories in DBpedia, and finally, the WCG was traversed via the shortest paths to the top-level categories.

6.1.2 Results and Discussion

Table 6.1 displays the number of posts by type found in the datasets. The column “unknown” refers to posts that were identified neither as a question nor as an answer.

¹<https://stackexchange.com/>

²<https://archive.org/details/stackexchange>

³<http://astronomy.stackexchange.com>

⁴<http://biology.stackexchange.com>

⁵<http://chemistry.stackexchange.com>

⁶<http://christianity.stackexchange.com>

⁷<http://history.stackexchange.com>

⁸<http://law.stackexchange.com>

⁹<http://mathoverflow.net>

¹⁰<http://music.stackexchange.com>

¹¹<https://philosophy.stackexchange.com>

¹²<https://sports.stackexchange.com>

Table 6.1: Distribution of post type in the stack exchange datasets along with the average text length

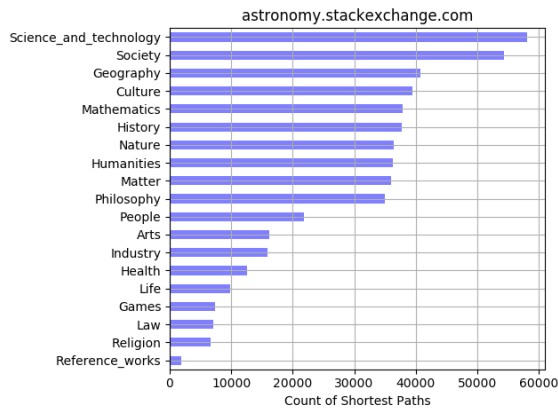
| Community | Questions | Answers | Unknown | Text length |
|--------------------------------|-----------|---------|---------|-----------------------|
| mathoverflow.net.count | 84,657 | 124,683 | 1,029 | 1100.06 \pm 1051.26 |
| chemistry.stackexchange.com | 23,074 | 26,997 | 646 | 1012.20 \pm 1130.93 |
| biology.stackexchange.com | 15,934 | 19009 | 1,068 | 1128.91 \pm 1227.67 |
| music.stackexchange.com | 11,101 | 29,980 | 770 | 992.32 \pm 979.54 |
| christianity.stackexchange.com | 9,267 | 22,043 | 1,446 | 1856.73 \pm 2074.10 |
| philosophy.stackexchange.com | 8,619 | 20,474 | 299 | 649.08 \pm 325.02 |
| history.stackexchange.com | 7,339 | 14,657 | 681 | 1355.08 \pm 1549.59 |
| law.stackexchange.com | 6,337 | 7,815 | 472 | 1197.79 \pm 1347.29 |
| astronomy.stackexchange.com | 5,019 | 7,383 | 437 | 1191.86 \pm 1368.82 |
| sports.stackexchange.com | 3,711 | 5,830 | 656 | 946, 36 \pm 1051.84 |

In figure 6.1, the results are displayed as the aggregated number of shortest paths based on the applied method from all documents of each dataset to each of the 19 top-level categories of Wikipedia. The distribution of the number of shortest paths among the 19 categories corresponds to the degree of relevance of each category in a given dataset. Thus, the category with the highest number of paths is the one that most contributes to the complete classification. An alternative visualization (as the percentage distribution) can be seen in Appendix B.

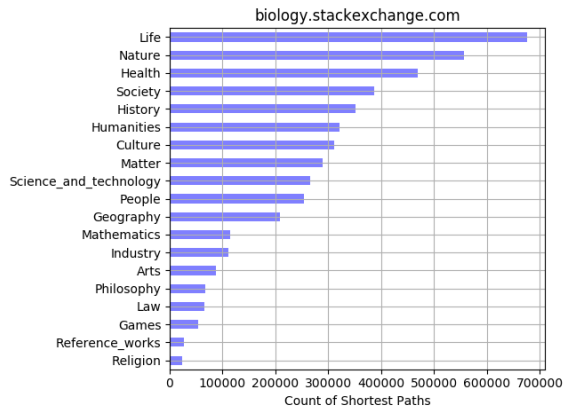
A high level of precision in the classification can be seen in the History (6.1e), Mathematics (6.1g) and Philosophy (6.1i) datasets, with the predominant category (the one with more shortest paths) corresponding directly to the topic of the community.

Although the names of the communities do not directly reflect a top-level category of Wikipedia, the results for the Astronomy (6.1a), Biology (6.1b), Chemistry (6.1c), Christianity (6.1d), and Sports (6.1j) datasets can also be considered accurate. As per the description, these communities were created to enable users to discuss the topics described by the categories whose paths are the most prominent.

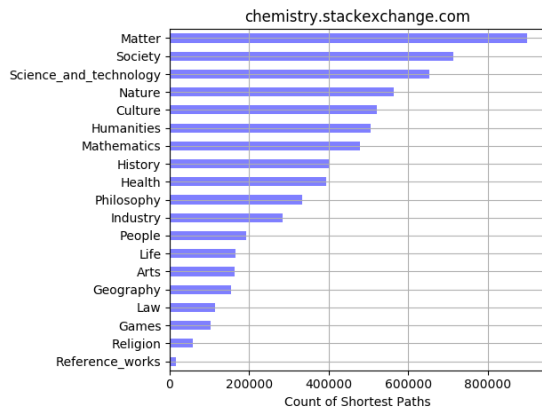
For Music (6.1h), the category with the highest number of shortest paths is Culture. This can be explained through understanding music as an essential aspect of any human society, and a form of (cultural) communication and expression.



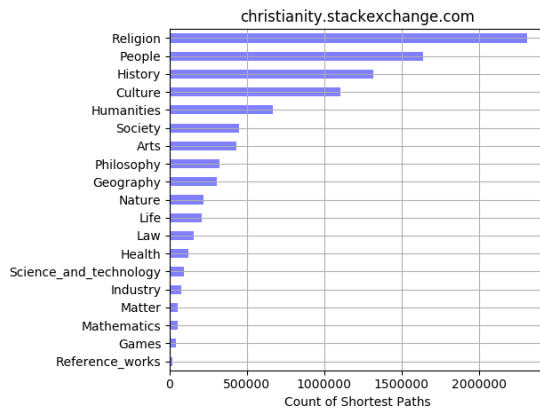
(a) Paths count for Astronomy



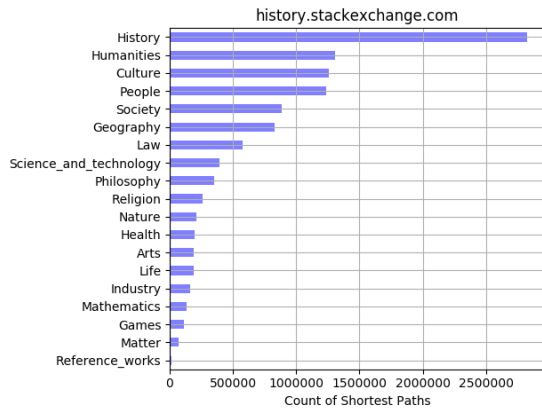
(b) Paths count for Biology



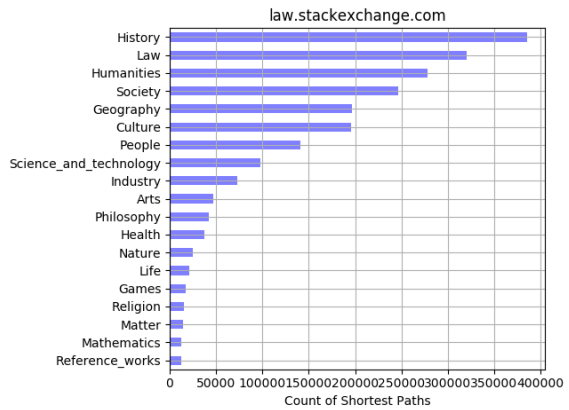
(c) Paths count for Chemistry



(d) Paths count for Christianity



(e) Paths count for History



(f) Paths count for Law

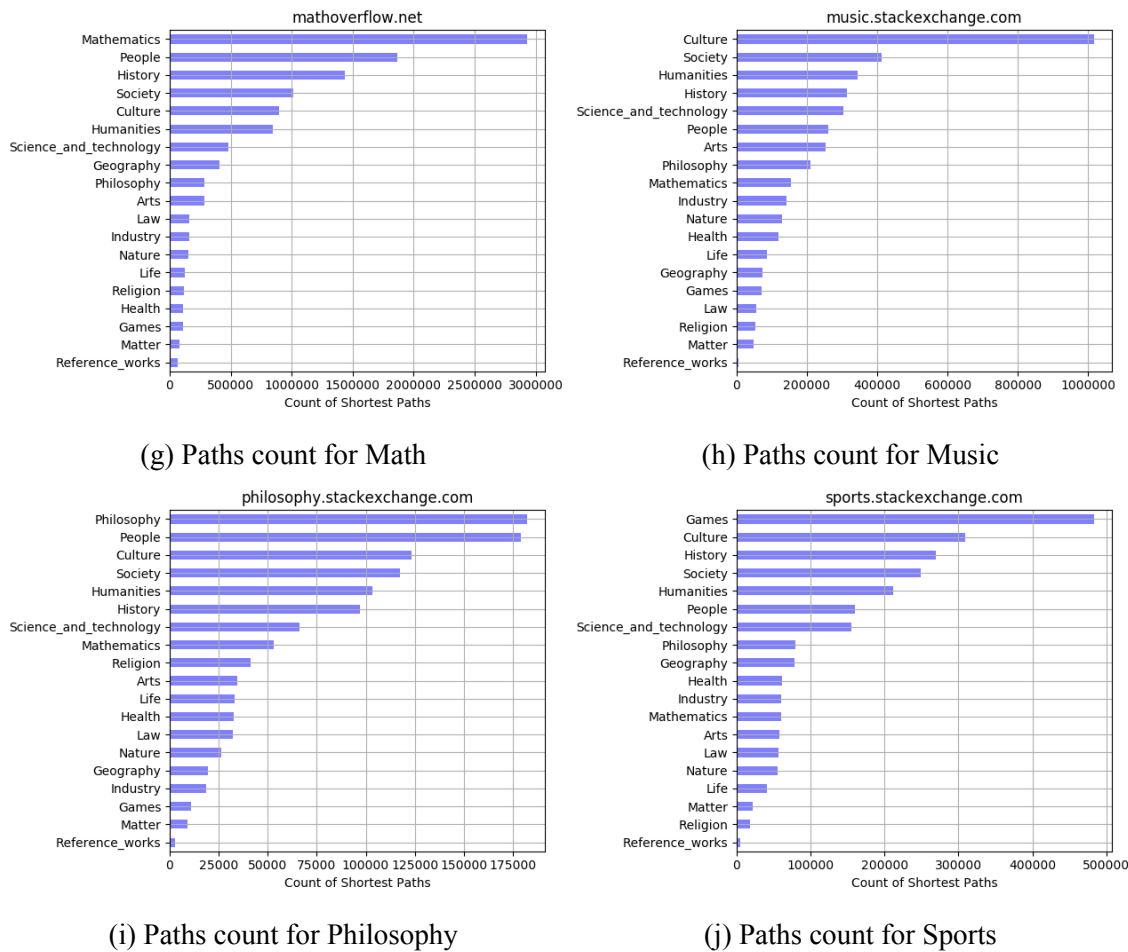


Figure 6.1: The number of shortest paths through the proposed method. The x-axis shows the number of paths found for each top-level category (displayed on the y-axis)

The Culture category has several subcategories representing different aspects of Music as a form of expression (e.g., Music by Genre, Music by Culture, Music in Culture), and the technical aspects of Music as science (e.g., Musical Composition, Music Terminology).

Another particular case is the result for Law (6.1f). As one of the 19 top-level is titled “Law”, it was expected that it would be the classification with the highest number of paths. It is however History, which appears with the highest degree, while Law appears second.

This can be explained as both concepts are closely related. New laws are passed in response to events occurring over the course of time - that is to say, history. One could argue that laws are a byproduct of history, but also that current laws will control future events (which, in turn, later become history themselves).

The example below of a genuine post (answer) extracted from the Law dataset illustrates how a topic related to Law is also connected to History. While there are the entities *Treaties*, *Court* and *Domestic Law* that are strongly related to the concept of Law, there are also the terms *Sovereign*, *Maastricht Treaty* and *Yugoslavia*, which are more strongly related to History.

“One of the powers that sovereign nations have is to make treaties with other sovereign nations, these can be bi-lateral (as in the example you cite) or multi-lateral (like the Maastricht Treaty that binds the EU together). Once a treaty is agreed and signed it needs to be ratified by each country which makes it part of the domestic law in that country: for your example, if India breaches the treaty it can be taken to court under the laws and in the courts of India or Pakistan(...). The world's newest nation is, I believe, South Sudan, and one that has recently vanished is Yugoslavia. Laws are not contracts: contracts require consent of the parties, among other things., laws don't, they are imposed irrespective of consent.”

As discussed in Chapter 3, the WCG presents the characteristics of a small-world network: a high level of connectivity between the nodes with relatively short paths. To some extent, the whole knowledge encoded in Wikipedia is interconnected. As a result, although the frequency distribution of categories is more densely clustered close to the y-axis, and the distribution curve tapers along the x-axis, it is important to note that there is at least a small percentage of relevance for each one of the 19 categories in the distribution for all ten datasets, as shown in figure 6.1.

6.2 Crowdsourcing study

Due to the accessibility of established micro-task crowd-sourcing platforms such as Amazon's Mechanical Turk¹³ and CrowdFlower¹⁴, researchers are actively turning toward

¹³www.mturk.com

¹⁴<http://crowdflower.com>

paid crowd-sourcing to solve data-centric tasks that require human input, such as building ground truths, validating results, and curating data [17].

In order to verify whether the classification generated by the applied method is coherent for real users, an experiment with human judges was conducted to ascertain the extent to which they agreed with the automatic classification result. To enable this comparison, Stack Exchange datasets were used in the experiment described in section 6.1.

6.2.1 Experimental Design

CrowdFlower was used to automatically allocate the available tasks to workers and to test them against known answers, namely the Gold Standard. Their performance on test questions indicates the extent to which the system trusts each worker – if they become untrustworthy, the user is removed from the task, and their work is discarded.

For this experiment, topics defined as a Main topic classifications on Wikipedia (19 categories by the date of data extraction) were also considered as top-level categories.

For each one of the ten communities, the experiment was run with 200 different random items extracted from the Stack Exchange dataset. Each worker was tasked with reading a randomly allocated text from the dataset, and ask to assert the extent to which they felt the text belonged to each one of the categories based on four options: i) not at all, ii) very little, iii) somewhat, or iv) to a great extent.

As described in [18], prior research publications have referred to the importance of task clarity tangentially and stressed the positive impact of task design, clear instructions and descriptions on the quality of crowd-sourced work. For this reason, a detailed guide was provided for the workers, to ensure they would understand how to perform the task, with examples of good and bad judgments and also with a description of each one of the 19 categories. To make sure the task is perfectly designed, CrowdFlower platform offers a consulting service where the top-rated workers evaluate and give feedback on the quality of a given task. An example of the feedback given by this consulting service can be seen in

figure C.2. The task description was adjusted according to the suggestions in the feedback by first providing more and contextualized examples, and second by explaining, for each test question, the reason why the alternatives were considered wrong or correct.

To alleviate the intensive task of judging for 19 Categories, the participants were asked to evaluate the top-3 categories with the highest percentage distribution and two other random categories. Figure C.1 shows a real example of a task delivered to workers for the dataset Biology. The categories of Life, Health, and Nature are the top-3 categories with the highest degree of membership (see figure 6.1b). The categories Religion and Games were randomly introduced into the survey.

Random categories were included to validate whether, in addition to agreeing with the categories that appear with the highest distribution in the classification, the workers would also agree with those that do not belong to the most prominent categories. Moreover, this mechanism reinforces the verification of the validity of the judgments, since random categories cannot have a distribution of responses similar to those in the top-3 categories.

To select the participants able to perform the task (and to eliminate those with low performance), a series of 30 test questions for each stack exchange dataset were created. After executing the study with users in the crowd-sourcing platform, an analysis was performed in order to verify the extent to which these users agreed with the classification generated by the approach used with the ten datasets extracted from Q&A communities.

6.2.2 Quality Control

When engaging a random collection of strangers to perform relevance evaluation, two primary concerns have risen: i) How to ensure the workers performing the evaluation will have the necessary skill or knowledge? ii) How to ensure that the workers will make an high-quality effort to do the work, rather than clicking randomly on the responses?

To address these questions, parameters for ensuring quality regarding the workers and the experiment were defined:

1. Participation was restricted to workers from English-speaking countries to ensure that they understood the task and instructions adequately.
2. Participation was restricted to Level-3 workers on CrowdFlower, meaning that only those who have completed over 100 test questions across hundreds of different types of tasks and have a near perfect overall accuracy were included. They are CrowdFlower's highest quality workers.
3. Each worker was restricted to a maximum of five judgments across all datasets, to minimize the number of workers trying to complete a disproportionately high number of tasks to maximize financial gain.
4. The value of 0.7 was asserted as the minimum level of agreement necessary for each row of evaluated text. In the case a row not reaching this value with the default number of three judgments, new judgments were requested until the level was reached. This value was chosen based on the suggestion of CrowdFlower platform. Values ranging from 0.71 to 0.80 were interpreted as substantial agreement [41].

Prior to running the experiment with all ten communities, the difference between the quality of judgments performed by elite workers and the judgments made by regular workers was evaluated. To perform this verification, the experiment was run using the Biology dataset with two different groups: one with regular workers exclusively, and the other with level-3 workers. Both were asked to judge the same set of 200 questions.

To compare the judgment made by the workers and the classification generated by the proposed method, the percentage of answers given in each of the categories of the scale (not at all, very little, somewhat and to a great extent) were aggregated for each of the evaluated texts. The precision, recall, and F-measure commonly used to verify the quality of IR techniques, including text classifiers [38] were then calculated.

Precision measures the number of times a category was correctly predicted by the proposed method (True Positive) divided by the number of times that category was predicted

in total (True Positive + False Positive). To maximize precision, the classifier must not fail to accurately classify the text entries in the dataset. Texts that should be assigned to one particular category according to the user study must be classified with the same category by the proposed automated approach. The main disadvantage of this metric is that it does not take into account the texts that should have been classified in a particular category, but were assigned to another one.

The recall bridges this gap by measuring the number of times a category has been assigned to a text by users (True Positive + False Negative), but the automatic classifier did not classify it correctly (False Negative). The disadvantage of this metric is that if the automated classifier did not classify any texts incorrectly, the recall would be maximum, even it was not efficient (because it also failed to sort correctly).

Measure F corresponds to the harmonic mean between precision and recall. With this information, the performance of the classifier can be asserted with an indicator only. The F-measure metric measures the efficiency of the classifier taking into account the error in both classes (True Positive and False Negative). It is necessary that the adjustment in both classes increase so that the metric increases. Considering that F-measure is an average, it gives a more accurate view of the efficiency of the classifier than just precision or recall.

6.2.3 Results and Discussion

1,265 unique workers participated in the final experiment, carried out between April and August 2018. The overall setup for the experiment is presented in table 6.2. A trusted judgment is an answer from a worker with an accuracy score higher than the minimum accuracy considered for this experiment (0.70), while an untrusted judgment is an answer from a worker whose accuracy score has fallen below this value.

Table 6.2: Overall setup for the experiment with crowd workers

| Community | Trusted Judgments | Untrusted Judgments | Average Judgments per Row | Average Trust of Workers | Unique Workers |
|-------------------|-------------------|---------------------|---------------------------|--------------------------|----------------|
| Astronomy | 640 | 36 | 3.2000±0.73 | 0.8651±0.061 | 128 |
| Biology (Elite) | 612 | 48 | 3.0600±0.57 | 0.8390±0.075 | 122 |
| Biology (Regular) | 1040 | 278 | 5.2000±1.78 | 0.8366±0.073 | 208 |
| Chemistry | 751 | 12 | 3.7550±1.22 | 0.8197±0.087 | 150 |
| Christianity | 628 | 16 | 3.1558±1.25 | 0.8843±0.073 | 125 |
| History | 602 | 76 | 3.0251±0.78 | 0.7997±0.070 | 120 |
| Law | 601 | 0 | 3.0050±0.66 | 0.8831±0.094 | 120 |
| Math | 630 | 32 | 3.1500±0.61 | 0.9306±0.078 | 126 |
| Music | 602 | 28 | 3.0100±1.10 | 0.9137±0.092 | 120 |
| Philosophy | 717 | 108 | 3.5850±0.83 | 0.8376±0.068 | 143 |
| Sports | 629 | 32 | 3.1608±0.12 | 0.8349±0.073 | 125 |

6.2.3.1 Elite workers vs Regular workers

A statistical test was applied to determine whether the results obtained in the answers given by the two groups can be considered significantly different from each other. The Wilcoxon-Mann-Whitney nonparametric statistical inference test was applied [12] to compare the mean of two independent samples, as it does not require normality and homoscedasticity in the series of values compared. The level of significance was set at 95%. The p-value evaluates the result of this statistical test. The lower the p-value, the higher the significance of the outcome. For a significance level of 95%, if the p-value is smaller than 0.05, the responses observed for the two groups can be considered as distinct.

A p-value of 0.267 (\gg 0.05) was obtained, meaning that there is no evidence of significant difference between the answers given by the two groups. Analyzing table 6.2, it is possible to see that regular users are less effective and thus more judgments are needed to achieve a reasonable level of agreement. Although the price paid to regular workers is lower by comparison to elite workers, a decision was made to admit only level-3 users.

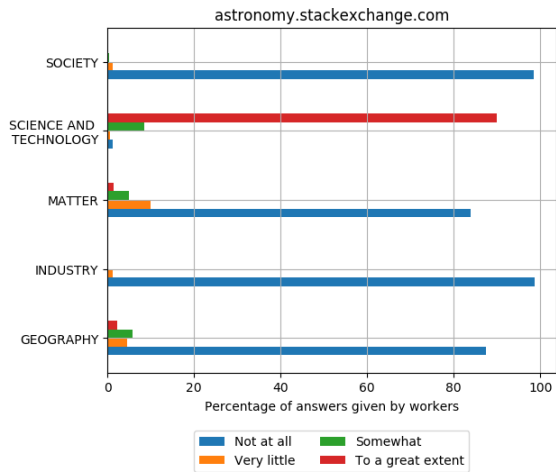
6.2.3.2 Human Judgment Analysis

Figure 6.2 shows the aggregated results of the participants' judgments for the categories in the Q&A community datasets. The vertical axis presents the five categories

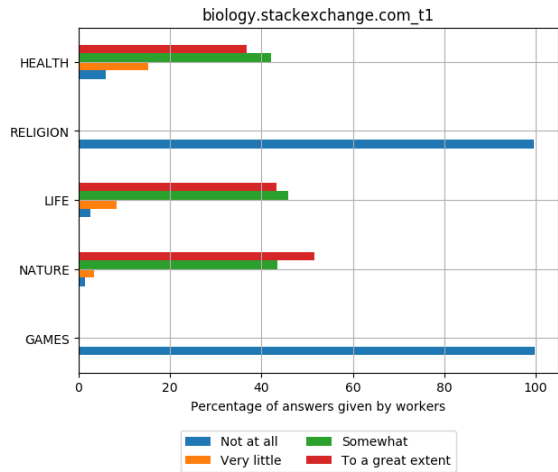
involved in each study, three of which correspond to the categories with the highest degree of membership obtained in the experiment described in section 6.1. The other two are randomly inserted categories.

In the Astronomy (figure B.1a), Biology (figures 6.2b and 6.2c), Chemistry (figure 6.2d), Christianity (figure 6.2e), Mathematics (figure 6.2h), Music (figure 6.2i) History (figure 6.2f), Philosophy (figure 6.2j) and Sports (figure 6.2k) communities, the category with the highest degree of relevance according to the proposed method were also the ones with the highest percentage of users asserting that they agreed with that category to a great extent. The most interesting findings regarding the comparison between the automatic classification and the answers given by users are highlighted below.

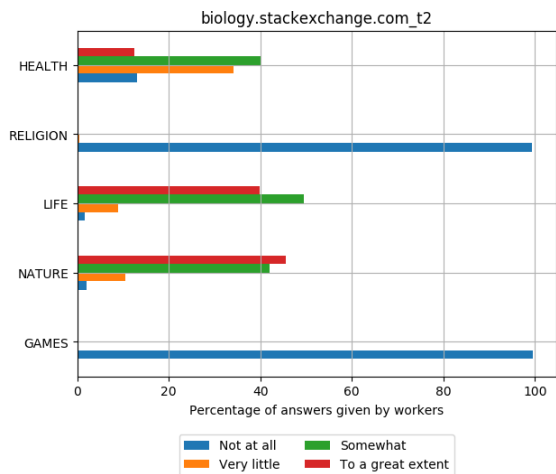
For the Law (figure 6.2g) category, while the proposed method identified History as the most relevant category in the dataset, the majority of users did not agree with this classification (97.70% not at all). The users identified Law as the most relevant category in the dataset (55.31% Somewhat and 42.95% To a great extent), whereas the proposed method suggested it as the second most relevant category. From this observation, it is possible to infer that the automatic extraction of categories from the entities named in the text can capture more subtleties of the information, while the users tend to perceive (in most cases) only the knowledge explicitly described in the text.



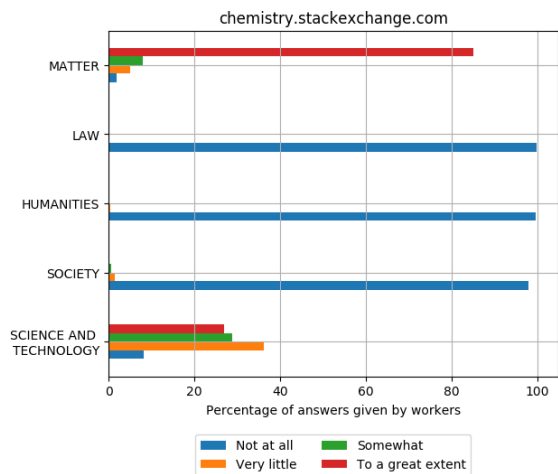
(a) Distribution of Answers for Astronomy



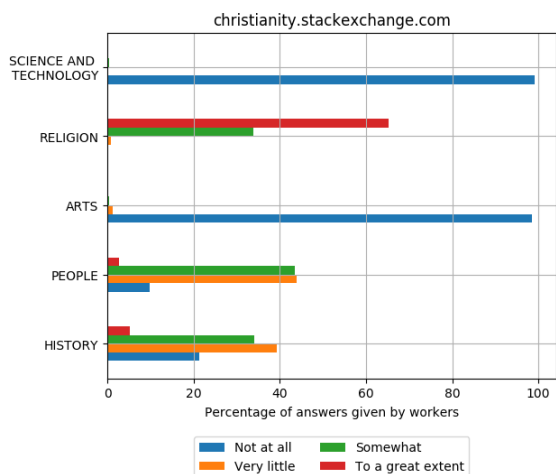
(b) Distribution of Answers for Biology (Regular)



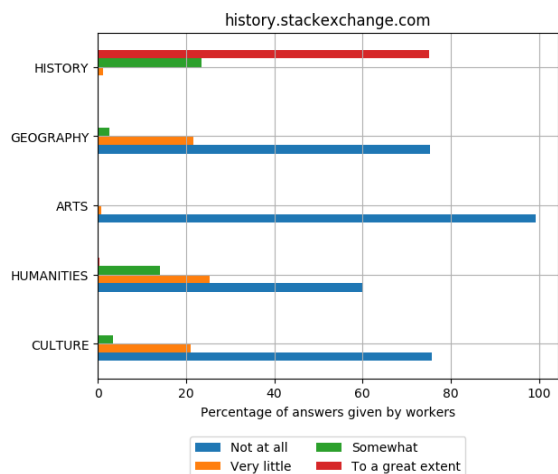
(c) Distribution of Answers for Biology (Elite)



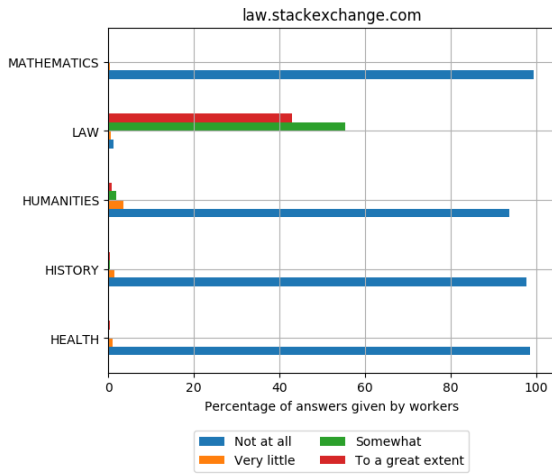
(d) Distribution of Answers for Chemistry



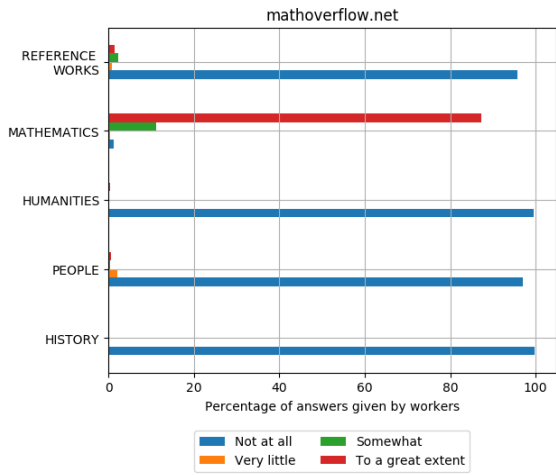
(e) Distribution of Answers for Christianity



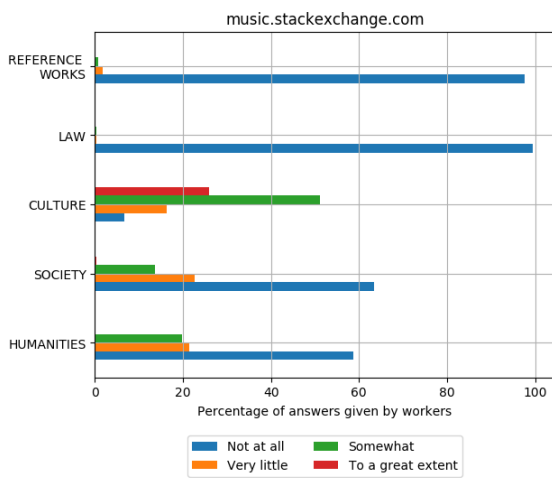
(f) Distribution of Answers for History



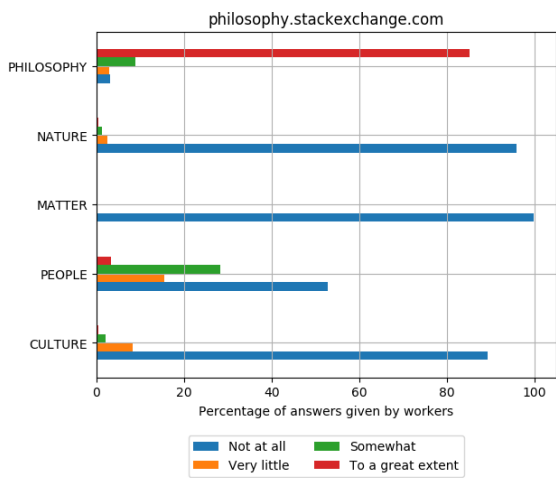
(g) Distribution of Answers for Law



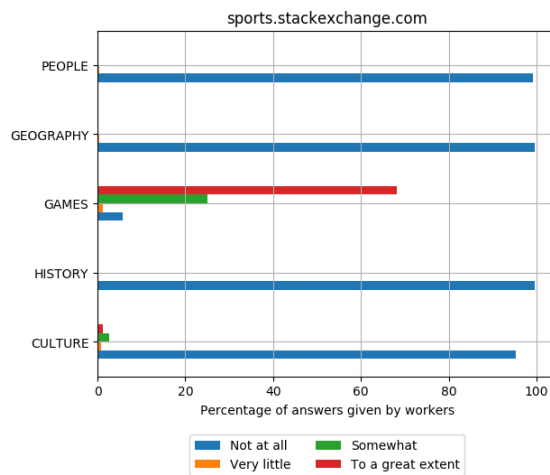
(h) Distribution of Answers for Math



(i) Distribution of Answers for Music



(j) Distribution of Answers for Philosophy



(k) Distribution of Answers for Sports

Figure 6.2: Percentage distribution of answers given by crowd contributors for each one of the ten communities evaluated

In the Astronomy (figure 6.2a) community, the three most relevant categories were Science and Technology, Society, and Geography respectively. The categories inserted randomly were Industry and Matter. For the most relevant category (Science and Technology), the vast majority of users agreed with the classification generated by the proposed method (89.86% To a great extent, 8.52% Somewhat, and 4.57% Very little).

The category Society was the second most relevant according to the automatic classification, but it was not identified by the users (87.51% Not at all). This can be explained by the fact that users identify categories superficially, based on their prior knowledge, but fail to recognize more tacit subjects that permeate the discussions. A good example that justifies the presence of the Society category in the automatic classification is the presence of several inquiries regarding the Geocentric Model, that is linked to the categories History of astrology, History of astronomy and Obsolete scientific theories, and indirectly connected to the category Society. However, users tend to answer that there is no relationship at all to the category Society in these discussions.

Although the category Matter was not one of the top 3 most relevant according to the proposed method's classification, it was identified as such by some users (9.88% Very little, 4.94% Somewhat and 1.35% To a great extent). The many discussions in the Astronomy community regarding the existence of carbon, water and other elements in the surface and the atmosphere of planets are the likely explanation for this phenomenon.

In the History community dataset (figure 6.2f), although the proposed method did not identify the category Geography as one of the top-3 most relevant categories (hence, it was randomly inserted in the experiment for this dataset), a considerable number of users asserted some degree of relevance to this category (21.70% Very little, 2.68% Somewhat and 0.24% To a great extent).

To extend the analysis, the precision, recall and F-measure were calculated for each of the texts judged by the workers and the automatic classification. Table 6.3 shows the summarized results. For this analysis, two levels of comparison were considered - they are identifiable in table 6.3 as L1 and L2. The first level (L1) considered the proposed

method to be correct when for a given text the category with the highest level of relevance was also identified by the user as related to the text to a great extent. The second level of comparison (L2) considered the proposed classification as correct when the user identified the text as related somewhat or to a great extent with the category suggested as the most relevant by the proposed method.

Table 6.3: Values of precision, recall and F-measure when comparing our classification and the judgments made by workers in the crowdsourcing study

| | L1 | | | L2 | | |
|------------------------------------|-----------|--------|-----------|-----------|--------|-----------|
| | Precision | Recall | F-Measure | Precision | Recall | F-Measure |
| Community | | | | | | |
| astronomy.stackexchange.com | 0.9423 | 0.2022 | 0.3330 | 0.9872 | 0.1935 | 0.3235 |
| biology.stackexchange.com | 0.4170 | 0.3311 | 0.3691 | 0.9064 | 0.3492 | 0.5041 |
| chemistry.stackexchange.com.csv | 0.8571 | 0.3111 | 0.4565 | 0.9524 | 0.3160 | 0.4746 |
| christianity.stackexchange.com.csv | 0.6815 | 0.5951 | 0.6353 | 0.9960 | 0.5731 | 0.7275 |
| history.stackexchange.com.csv | 0.7528 | 0.6526 | 0.6991 | 0.9925 | 0.6559 | 0.7899 |
| law.stackexchange.com.csv | 0.0149 | 0.6667 | 0.0292 | 0.0299 | 0.6667 | 0.0571 |
| mathoverflow.net.csv | 0.8896 | 0.6361 | 0.7418 | 0.9955 | 0.6314 | 0.7727 |
| music.stackexchange.com.csv | 0.2760 | 0.8019 | 0.4106 | 0.8019 | 0.7816 | 0.7917 |
| philosophy.stackexchange.com.csv | 0.9081 | 0.3889 | 0.5446 | 0.9622 | 0.3732 | 0.5378 |
| sports.stackexchange.com.csv | 0.7372 | 0.5479 | 0.6286 | 0.9805 | 0.5331 | 0.6907 |

For the majority of datasets, high precision (> 0.70) and a moderate recall (> 0.50) were obtained, meaning that the proposed method identifies one category as the most relevant when the users pointed it as related to a great extent with the text. The users did however sometimes identify a category that was not the most relevant according to the proposed method as the one with highest relation to the text.

In the Biology and Music datasets, when considering L1 as the basis for the comparison, a low value for precision was obtained. This is mainly because there is a better distribution of answers as “to a great extent” along the 3 most relevant categories pointed out by the proposed method than in the other datasets (figures 6.2c, 6.2c and 6.2i). If we consider L2 as the basis of the comparison instead, the precision increases significantly

for both datasets.

The Law dataset is the only one to show low values for precision for both L1 and L2. This is because while the proposed method identified the category History as the most relevant, the majority of users identified only the category Law as being related to the texts to a great extent.

7. Related Works

The knowledge encoded in Wikipedia has been used by many researchers as a tool for performing several tasks, including text categorization [15], co-reference resolution [67], predicting document topics [60], automatic word sense disambiguation [46], searching synonyms [32] and computing semantic relatedness [55], [16], [49].

The method proposed in this thesis aims to categorize text-based resources based on the Named Entities found in the text and assert its relation to a set of predefined categories in Wikipedia in a way that users can understand and make use of. The Wikipedia category structure was represented as a graph and used to determine the categorization based on the shortest paths between the categories associated with the entities and a set of more generic predefined Wikipedia categories. Classifiers can be created in many forms, and to focus on different features. There are a number of existing projects that share the goal of creating text classifiers from a knowledge base:

- Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge [15]
- What’s in Wikipedia?: mapping topics and conflict using socially annotated category structure [30]
- Identifying document topics using the Wikipedia category network [60]
- Bringing Bag-of-phrases to ODP-based Text Classification [63]

- Toward Robust Classification using the Open Directory Project [23]
- A Method for Automated Document Classification Using Wikipedia-Derived Weighted Keywords [6]
- Classification of Comments by Tree Kernels Using the Hierarchy of Wikipedia for Tree Structures [68]
- Wikitop: Using Wikipedia Category Network to Generate Topic Trees [33]

Gabrilovich and Shaul Markovitch [15] described a method for finding the Wikipedia article most similar to a given document, and extends that document's BOW representation with the words occurring in the Wikipedia article. This approach provides a greater number of topic-specific words to the documents, which makes it easier to classify them with standard text classification techniques. The idea behind this approach is similar to the one outlined in this thesis regarding the attempt to provide richer semantics to the representation of documents than could be achieved using the BOW approach. However, their approach is very complex and yet fails to add semantic context to the representation.

The method proposed by Kittur and Chi [30] is similar to the approach outlined in this thesis. Their goal was to automatically assign a Wikipedia article to a set of what they call macro-categories (a subset of Wikipedia categories that is at the top of the hierarchy). The main difference is that their approach is limited to articles inside Wikipedia, and cannot be generalized for other text-based resources – this limitation is, however, addressed in the research outlined in this thesis. Kittur and Chi [30] evaluated the approach by comparing the attributions made by their method with the judgment of human raters. Although they present a moderate positive correlation between the method and the judgments, the experiment was realized on a small scale by comparison to the one carried out in this thesis.

A similar approach was applied by Schönhofen [60] to determine whether documents could be categorized by exploring features from Wikipedia. They validated their method first by predicting categories of the Wikipedia articles themselves, and then by classifying documents of an external dataset based on their Wikipedia categories. The main difference

between this research and the one reported on in this thesis is that in the latter, the entire category structure has been taken into account, while the approach described in [60] looked exclusively at categories retrieved from the matched Wikipedia article titles (hence, their approach was more limited). Leveraging the hierarchical structure of the categories, as was completed in this thesis, was allocated to future work by Schönhofen.

Ha et al. [23] addressed the problem of sparsity in the Open Directory Project categorization structure by testing several approaches for text classification. They demonstrated that training data expansion is one of the promising directions to deal with the sparse characteristic of the ODP dataset. One of the interesting findings of this work is that distance-based weighting had a better result over the other methods tested. The approach was only evaluated in pages manually classified in the ODP.

Shin et al. [6] proposed a method for overcoming the limitations of BOW by representing the texts with a group of phrases rather than words alone. They employed a syntactic tree to extract phrases from Open Project Directory and applied a phrase selection method to alleviate the high dimensionality problem of bag-of-phrases. Although the approach proposed by them shares the goal of providing more semantic features for document representation in the classification task, it makes use of a knowledge base that has been built by experts, and discontinued in 2017. Hence, the approach outlined in this thesis - which leverages the knowledge contained in Wikipedia - is broader and more flexible to change.

Biuk-Aghai and Ng [6] also presented a method for the automatic classification of scientific articles based on the categories of Wikipedia. They too take advantage of the category graph to generate the classification. The first difference is that while the method utilised in the context of this thesis extracts concepts from text based on the recognition of named entities, [6] uses a statistical approach to extract keywords considered relevant. The second difference is that to aggregate the categories that represent the document, unlike the method described in this thesis, which uses the shortest paths towards the main topics, [6] used a measure of semantic similarity to find the most related categories. The quality of their method for the classification was only evaluated manually by the authors.

Takeda et al. [68] described a method for the classification of tweets in a system for IR in the context of tourism information on social media. They propose a technique for classification that utilizes tree kernels (topic trees) created from categories extracted from Wikipedia. Although the approach presented in [68] is similar to the method used here, their research did not take full advantage of the rich structure of categories since they only considered paths that are three levels deep. Furthermore, they transformed the graph into a tree by taking into account only one shortest path from each category to the top, while the analysis described in this thesis included the topology of the graph (supporting the decision to keep it as a small-world network).

Kumar, Rengarajan Annie [33] described Wikitop, a method for automatically generating topic trees from the text by performing hierarchical classifications using the Wikipedia Category Structure. The major difference between their approach and the one outlined here is that, while the proposed method assigns flat categories with degrees of pertinence, Wikitop assigns a hierarchical classification (topic trees). Furthermore, the technique has only been tested on a small scale and only with Wikipedia's articles.

8. Final Remarks, Limitations and Future Works

This, the final chapter of the thesis, provides the conclusion for the outlined research. It begins with the final remarks on the project, before mentioning some limitations and desired future work that might improve the classification results of the proposed method.

8.1 Final Remarks

This thesis presented the construction of a method for the automatic classification of textual resources on the Web, which exploits the collective knowledge of the Wikipedia contributors rather than the effort of domain experts.

The central motivation to reduce the need for experts to mediate the classification process is the fact that the amount of information generated on the Web grows as more people use the platform. As a consequence, most efforts to classify and organize documents manually on the Web have proven to be unviable and have become extinct. Simultaneously, there is a great movement of people who come together to create and organize content on the Web collaboratively.

This form of classification has the advantage of being dynamic and representing the way people organize the areas of knowledge. It is robust to any change in facts, people, places, etc., and can be quickly edited by contributors.

In this context, the decision was made to develop a method for representing and classi-

fyng Web documents based on the top-level categories of Wikipedia, since these are easy for regular users to understand, and can be easily modified to serve specific purposes.

8.2 Contributions

Given the complex structure of the Wikipedia category system, the decision was made to represent it as a graph whose nodes represent the categories and the edges represent the “is-sub-category-of” relationships. For presenting a complex structure with several categories, many links, and cycles, it was necessary to perform an analysis of the topology of the category graph of Wikipedia, in order to verify whether this structure was suitable for the application of the proposed method.

The analysis leads to the conclusion that the Wikipedia Category structure is similar to other semantic networks often used for NLP applications. It was verified that the WCG presents a small-world and scale-free behaviors. This finding supports the Wikipedia categorization scheme not only for the developed classification method but also for other NLP and IR applications.

A new method for extracting features, representing and classifying documents in a three-steps processing chain was proposed. In this approach, the named entities present in the text are recognized, and the categories of those entities and aggregate the representation are extracted into a predetermined set of topics within Wikipedia categories. The main advantage of this approach is that it captures semantic information of texts, even if they are short. In this regard, it is different from the traditional approach that uses only word frequency without considering context.

As one of the goal goals of this thesis comprises of allowing users of IR applications to understand and make use of the classification, an experiment to verify that real users do indeed agree with the classification generated by the approach was carried out. The study involved 1,265 users of a crowd-sourcing platform and 2,000 different texts extracted from ten Q&A communities. The results showed that for most cases, users agree to a great

extent with the classification generated by the developed method. Although the users did not agree with the automatic classification in some cases, it was observed that the classification made sense concerning the content of the texts. However, some subtleties regarding text details and transversal topics were not captured by users, who tended to make judgements based on the general context of the given text.

A secondary technical contribution is TagTheWeb¹, a public, documented² and open-source API capable of receiving any textual resource and processing each of the three phases described in the proposed approach.

The task of extracting information from Wikipedia and representing it as a graph posed a substantial challenge for the course of this dissertation. Calculating the metrics needed to evaluate the topology of the WCG, given its dimension, was computational costly and time-consuming. As a technical legacy, there are the WCG snapshot from October of 2016 filtered and represented in Neo4J³ and graph-tools⁴ and a dataset⁵ containing all nodes of the WCG and the measures of centrality, indegree, outdegree, clustering coefficient, and PageRank, that can be used to alleviate this task in future works.

8.3 Limitations

Despite an optimistic initial result, the reported research has some limitations. Because it is based on the extraction of named entities, if no entity is recognized in the text, classification is not possible. Moreover, if only one entity existed in the text, it could be said that the proposed method generates a classification for the entity, not for the context presented in the document.

Concerning the recognition of entities, the designed approach is dependent on a tool to recognize and link the entity to DBpedia. However, if entities are incorrectly recognized

¹<http://ww.tagtheweb.com.br>

²<http://documenter.getpostman.com/view/1071275/tagtheweb/77bC7K>

³<http://neo4j.com>

⁴<http://graph-tool.skewed.de>

⁵<http://github.com/jerrylewisbh/TagTheWeb>

(especially when disambiguation fails), the classification may be incorrect. This problem is intensified in the case of short texts.

Another concern is regarding the topics distribution in Wikipedia. For example, that are many more categories associated with History than with Games. This could result in a biased or unbalanced categorization.

An additional limit of this research is the dimensionality of the document representation. Limiting the classification to the 19 top-level categories of Wikipedia made it possible to run the experiments, but, as a consequence, all results and observation are directly related to the content expressed in these categories.

Regarding the evaluation, only 200 texts from each one of the ten Q&A communities were used, due to time and cost constraints. The results and conclusions are based on the observations on these ten communities. Further analyzes in a larger number of more diverse communities and in other contexts are necessary to produce more reliable results regarding the quality of the classifier.

8.4 Future Works

An expansion of the research presented in this thesis would be to study the use of other sets of categories as the main topics in the representation and classification of documents, according to specific contexts of use. This study was limited to Wikipedia's broader categories, but, given the structure of the graph, the developed method could be applied to any subset of categories at any level of depth.

Considering that Wikipedia and DBpedia are available in a wide variety of languages, it is also desirable to expand the experiments to verify the quality of the classification in different Wikipedias, since not all versions are as broad and complete as Wikipedia in English.

Studying methods for cleaning, pruning and organizing this structure is an opportunity

of future work that could reduce the complexity of the graph and improve the results. Since the concept “subcategory of” is not well defined and there is no policy to guide the users who contribute, the category graph is far from perfect: there are many duplications, misplaced categories, excessive fragmentation and cycles in the way. This improvement would bring up a more clear and better distribution of categories.

A.Representing the Underlying Structure of the WCG

In the scope of this thesis, finding the paths between entities and a set of top categories requires that the structure of Wikipedia Category be represented so that computer programs can navigate on it. Therefore, the proposed method needed a way of representing the available information about the Wikipedia structure as a graph. The WCG consists of Wikipedia pages with the “Category:” prefix such as “Category:Law.” The graph is extracted by finding links between category pages. In other words, a category page is linked to another category page that is broader in scope.

To perform this task, a dump of the files `enwiki-latest-page.sql`, `enwiki-latest-category.sql`, and `enwiki-latest-categorylinks.sql`, was obtained in October 2016 ¹.

These files consist the structure of Wikipedia represented in a MySQL relational database.

The file `enwiki-latest-categorylinks.sql.gz` contains the information needed to extract the categories-categories and categories-articles relations. The information is represented in the file as INSERT statements where all entries are on the form:

(cl_from,cl_to,cl_sortkey,cl_timestamp,cl_sortkey_prefix,cl_collation,cl_type).

¹<https://dumps.wikimedia.org/enwiki/20161020/>

| Field | Description |
|-------------------|--|
| cl_from | Stores the page.page_id of the article where the link was placed |
| cl_to | Stores the name (excluding namespace prefix) of the desired category. Spaces are replaced by underscores (_) |
| cl_sortkey | Stores the title by which the page should be sorted in a category list. |
| cl_timestamp | Stores the time at which that link was last updated in the table. |
| cl_sortkey_prefix | either an empty string if a page is using the default sortkey or a human readable version of cl_sortkey. |
| cl_collation | What collation is in use. |
| cl_type | What type of article is this (file, subcat (subcategory) or page (normal page)). |

Table A.1: Description of fields in INSERT statements for the table categorylinks.

Considering the goal of this process is to represent the underlying Wikipedia category structure as a graph, Neo4J², was chosen because it is a graph-based database that provides a free version and proper documentation, as well as an active community.

Even though only the category graph is needed, it was necessary to extract the page file as well, since the identifier of the categories refers not to the categories themselves, but to the page about the categories. For this step, a parser was developed in python. First, the file containing all the pages of Wikipedia is scanned and, through a regular expression, the data is filtered.

The output is a CSV file with three fields: page identifier, page title, and namespace. Namespaces are a Wikipedia internal classification system to identify what a page refers to, such as an article, a category, user page, or other classifications. Articles are identified with namespace 0 and categories with namespace 14, for example. The second step is to go through the file that corresponds to the categories of Wikipedia. The output of this step is a CSV file with two properties: page identifier and category title. From this file, it is possible to link the categories and the corresponding pages. The third and final step of this step is the extraction of links between categories. The data is extracted from the file of the enwiki-latest-categorylinks.sql file, and the output is a CSV file with two properties: page identifier for the source category and page identifier for the destination category.

Administrative categories are used for Wikipedia internal organization and maintain-

²<https://neo4j.com/>

ing. They were ignored during processing and are not in the final files. Examples of these categories include, but are not limited to, categories that begin with *Articles_needing_* and *WikiProject_*, named hidden categories.

These categories provide a tool for grouping pages with features in common so that publishers can quickly identify where improvements are needed. Although these categories are essential for the maintenance and administration of the encyclopedia, they are not relevant to end users and were removed in the context of the work for two reasons: 1) They add complexity to the structure of the graph 2) They do not semantically describe articles associated to them and therefore do not add relevant information to the paths.

The ideal paths between a set of categories have no hidden categories. Thus, the hidden categories must be removed in such a way that the information is not lost because a hidden category can be a subcategory of a visible category or they can have visible categories as their subcategories. Hidden categories were removed and the paths were reconstructed the paths by linking the child nodes of the hidden category to its parent.

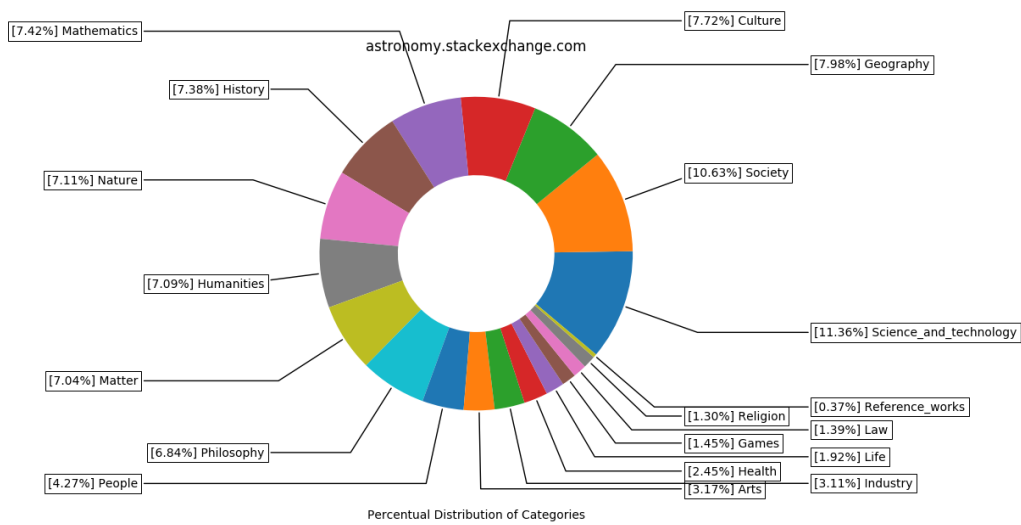
After processing the Wikipedia dump files, it is possible to generate the complete graph of categories. First, all the pages referring to categories are imported in Neo4J, and then the relations between them are imported. The generated graph has only a Category concept with the `categoryName` and `categoryID` properties and a `SUBCATEGORY_OF` relationship type, with no properties.

The example below illustrates how IR works in a Neo4J graph.

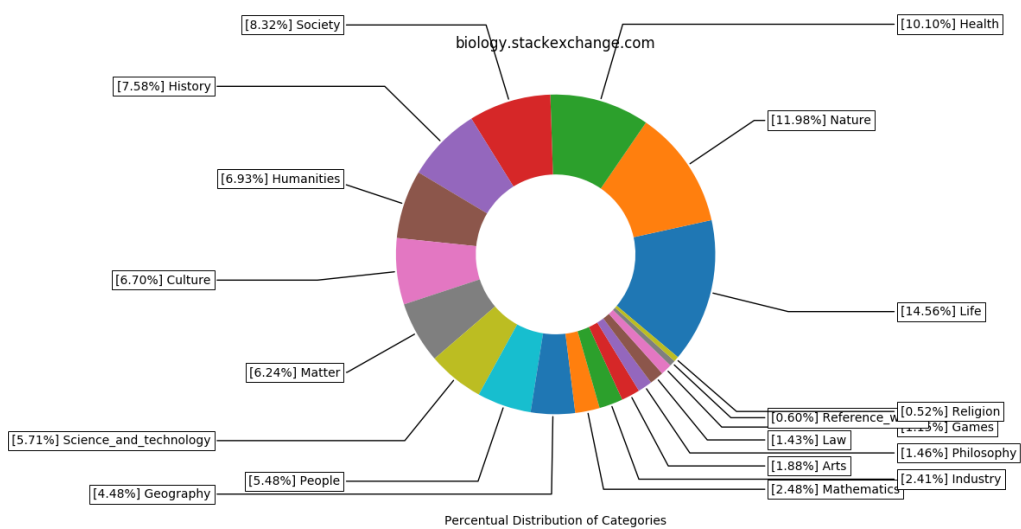
The example returns all nodes of type `Category` that are connected to another node of type `Category` that has the property `CategoryName` with the value equals to `Carnivores` for the `SUBCATEGORY_OF` property. In this case, all subcategories of `Carnivores` are retrieved.

```
MATCH (a: Category) - [r: SUBCATEGORY_OF] ->
(b: Category {categoryName: ""}) RETURN a
```

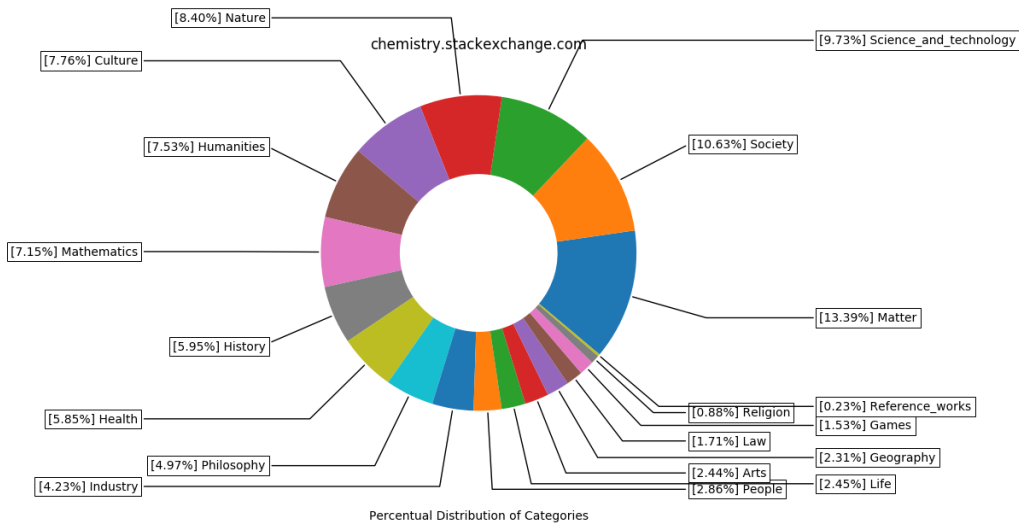
B. Percentage Distribution of Categories



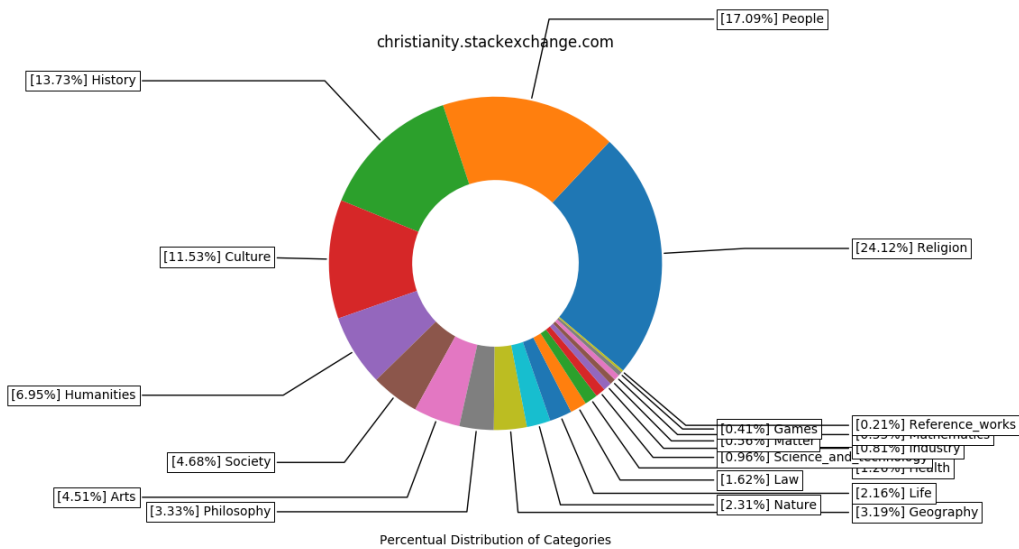
(a) Percentage distribution of categories for Astronomy



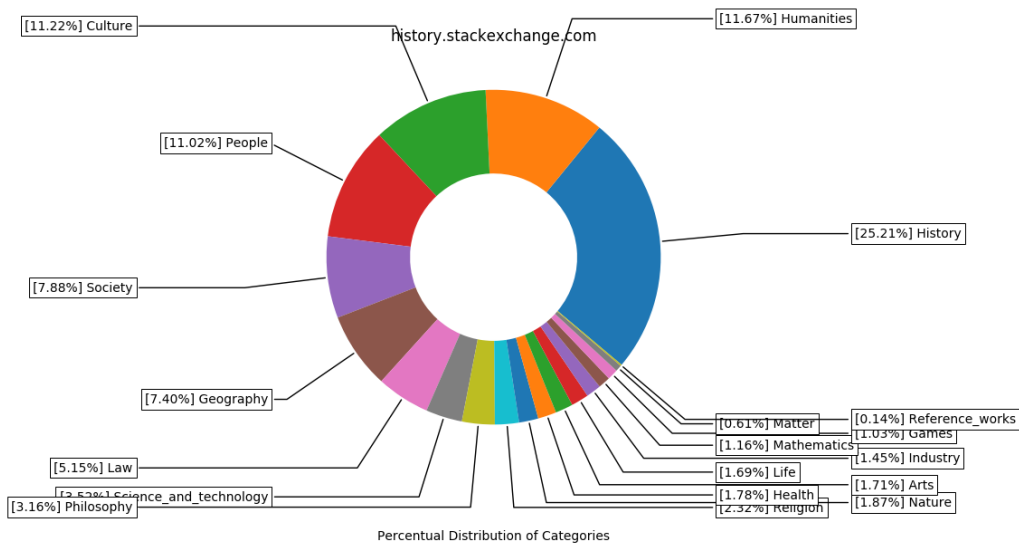
(b) Percentage distribution of categories for Biology



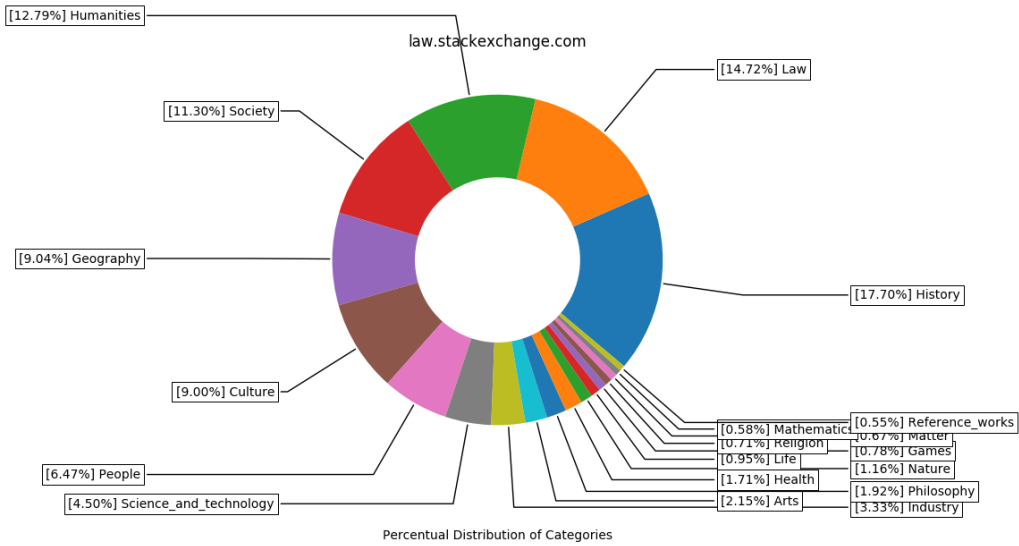
(c) Percentage distribution of categories for Chemistry



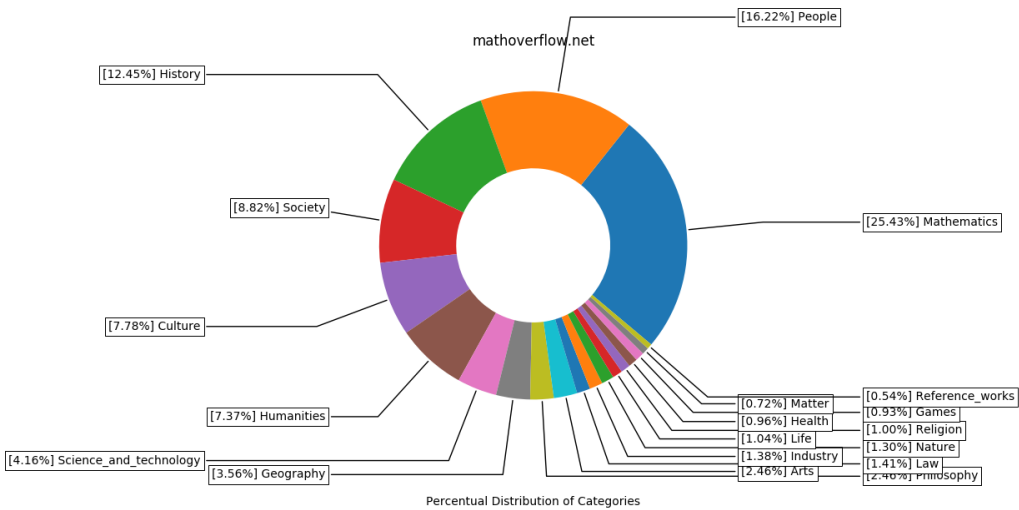
(d) Percentage distribution of categories Christianity



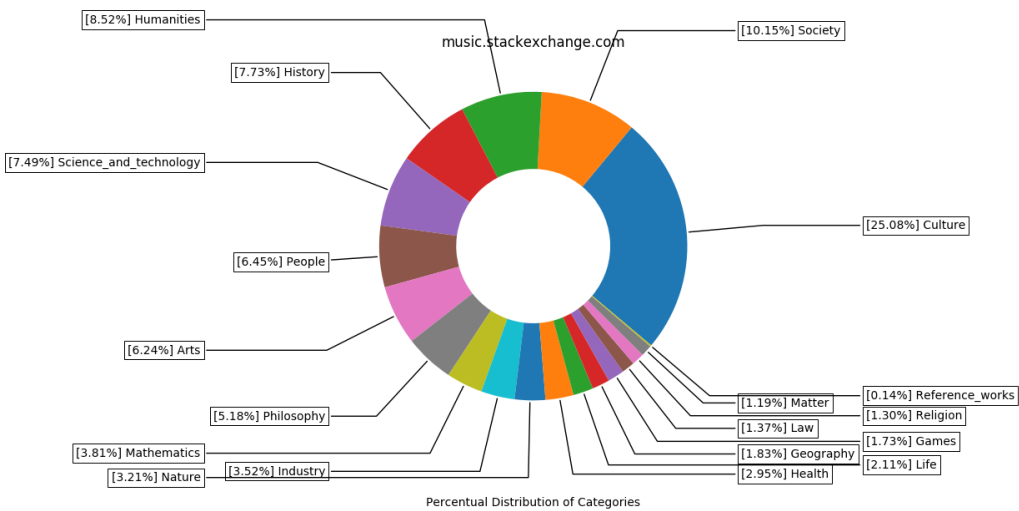
(e) Percentage distribution of categories for History



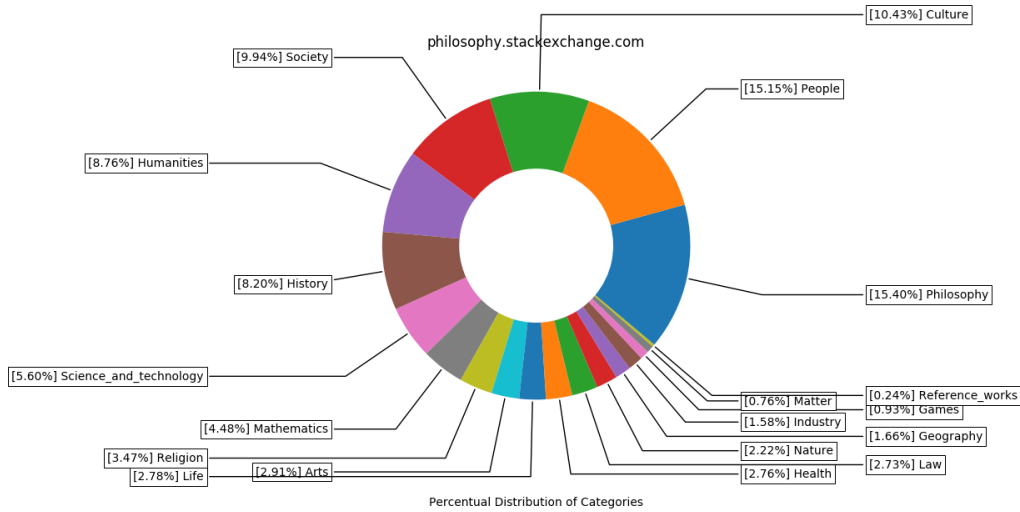
(f) Percentage distribution of categories for Law



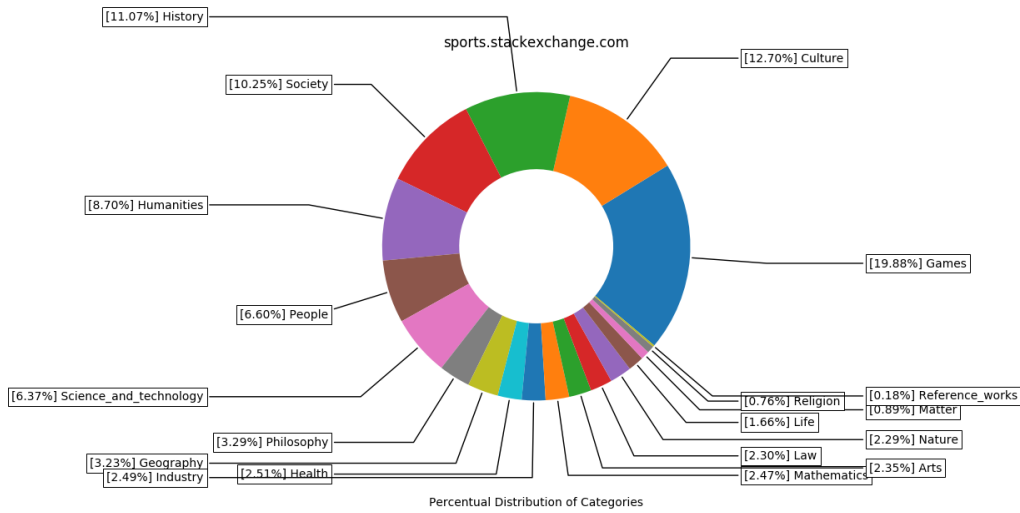
(g) Percentage distribution of categories for Math



(h) Percentage distribution of categories for Music



(i) Percentage distribution of categories for Philosophy



(j) Percentage distribution of categories for Sports

Figure B.1: Percentage distribution of categories for each of the communities evaluated according to the proposed method.

C.Crowdsourcing Experiment Details

In prokaryotic translation- how critical for efficient translation is the location of the ribosome binding site- relative to the start codon?

Ideally- it is supposed to be -7b away from the start. How about if it is -9 bases away or even more? Will this have an observable effect on translation?

LIFE: To what extent is this topic related to the text? (required)

- To a great extent
- Somewhat
- Very little
- Not at all

HEALTH: To what extent is this topic related to the text (required)

- To a great extent
- Somehow
- Very Little
- Not at all

NATURE: To what extent is this topic related to the text? (required)

- To a great extent
- Somewhat
- Very little
- Not at all

RELIGION: To what extent is this topic related to the text? (required)

- To a great extent
- Somewhat
- Very little
- Not at all

GAMES: To what extent is this topic related to the text? (required)

- To a great extent
- Somewhat
- Very little
- Not at all

Figure C.1: Example of task delivered to contributors in the dataset Biology

Feedback for JOB ID 1028469 'Text Classification [G2]'

Hello

Easy to understand instructions

Definitions for terms are well defined

Given example is good with analyzation

Terms used in the real tests are different to the given terms in Rules

Need definitions for the given terms in the real tests

Validations are working fine

Corrections have no explanations at all

<https://s3.amazonaws.com/uploads.hipchat.com/32366/5107918/ZbPSZfTxZo<fB9e/lx%20job.jpg>"

The corrections supposed to have more explanation as in the example given
Task author, please address the following 2 things to make this task more perfect:

1- Definitions for the terms/words in the real tests

2- Explanation for the corrections in more detailed way

If the above 2 are addressed, this task will be very good to work

That's it from me

Figure C.2: Example of feedback given by the consulting service provided by Crowd-Flower platform

D.Wikipedia Category Graph - Nodes Dataset

Table D.1: A sample including the 50 first rows of the dataset generated from the Wikipedia Category Graph. Each row represent one node of the graph along with its Degree, Clustering Coefficient, and Centrality information

| CategoryName | Degree | Outdegree | Indegree | Clustering Coefficient | Betweenness Centrality | PageRank |
|-------------------------|--------|-----------|----------|------------------------|------------------------|------------------------|
| Futurama | 15 | 11 | 4 | 0.00909090909090909 | 5,89E+08 | 4,27E+08 |
| World_War_II | 52 | 15 | 37 | 0.047619047619047616 | 2,72E+11 | 1,05E+11 |
| Programming_languages | 56 | 5 | 51 | 0.15 | 9,48E+08 | 8,17E+08 |
| Professional_wrestling | 27 | 4 | 23 | 0.0 | 2,02E+10 | 8,64E+09 |
| Algebra | 16 | 1 | 15 | 0.0 | 1,26E+09 | 7,42E+10 |
| Anime | 34 | 8 | 26 | 0.05357142857142857 | 1,66E+09 | 4,01E+09 |
| Abstract_algebra | 32 | 3 | 29 | 0.16666666666666666 | 7,26E+07 | 3,98E+08 |
| Mathematics | 29 | 7 | 22 | 0.14285714285714285 | 1,54E+07 | 1,33E+11 |
| Linear_algebra | 18 | 2 | 16 | 0.0 | 6,04E+08 | 1,25E+11 |
| Calculus | 11 | 3 | 8 | 0.5 | 4,42E+07 | 9,90E+08 |
| Monarchs | 36 | 6 | 30 | 0.16666666666666666 | 8,12E+09 | 4,73E+08 |
| British_monarchs | 16 | 7 | 9 | 0.09523809523809523 | 2,43E+08 | 6,87E+08 |
| Star_Trek | 28 | 9 | 19 | 0.027777777777777776 | 2,72E+09 | 1,57E+10 |
| People | 360 | 3 | 357 | 0.3333333333333333 | 9,76E+08 | 0.00032288729268168 |
| Popes | 47 | 18 | 29 | 0.0196078431372549 | 5,80E+08 | 2,59E+10 |
| Desserts | 23 | 2 | 21 | 0.5 | 4,16E+07 | 1,96E+10 |
| Fruit | 28 | 4 | 24 | 0.25 | 3,34E+08 | 2,71E+09 |
| Lists | 33 | 4 | 29 | 0.16666666666666666 | 1,92E+11 | 6,54E+09 |
| Computer_science | 21 | 7 | 14 | 0.2619047619047619 | 7,98E+09 | 6,61E+09 |
| The_Simpsons | 18 | 9 | 9 | 0.027777777777777776 | 1,69E+09 | 7,97E+08 |
| Algorithms | 58 | 6 | 52 | 0.06666666666666667 | 9,51E+08 | 4,56E+09 |
| Data_structures | 24 | 6 | 18 | 0.13333333333333333 | 9,86E+07 | 1,25E+09 |
| Monty_Python | 15 | 4 | 11 | 0.0 | 2,56E+07 | 6,11E+08 |
| Middle-earth_places | 0 | 0 | 0 | 0.0 | 0.0 | 1,02E+09 |
| Middle-earth_characters | 25 | 7 | 18 | 0.023809523809523808 | 9,36E+07 | 7,91E+08 |
| Middle-earth | 24 | 5 | 19 | 0.05 | 1,50E+09 | 1,23E+10 |
| Science | 38 | 2 | 36 | 0.0 | 3,22E+09 | 0.00010830385134322265 |

Table D.1 continued from previous page

| CategoryName | Degree | Outdegree | Indegree | Clustering Coefficient | Betweenness Centrality | PageRank |
|-------------------------|--------|-----------|----------|------------------------|------------------------|------------------------|
| Chemistry | 74 | 2 | 72 | 0.5 | 3,69E+09 | 1,93E+11 |
| Middle-earth_Valar | 5 | 5 | 0 | 0.05 | 0.0 | 1,02E+09 |
| Middle-earth_languages | 6 | 4 | 2 | 0.08333333333333333 | 9,53E+06 | 2,18E+09 |
| Middle-earth_books | 14 | 8 | 6 | 0.0 | 2,79E+08 | 4,52E+09 |
| Vietnam_War | 66 | 33 | 33 | 0.013257575757575758 | 6,55E+08 | 1,78E+10 |
| Middle-earth_Maiar | 7 | 6 | 1 | 0.03333333333333333 | 2,18E+06 | 1,31E+08 |
| Middle-earth_Elves | 8 | 4 | 4 | 0.08333333333333333 | 1,87E+07 | 3,45E+08 |
| Countries | 29 | 7 | 22 | 0.11904761904761904 | 7,88E+10 | 8,57E+10 |
| Middle-earth_Rohirrim | 3 | 2 | 1 | 0.0 | 8,29E+05 | 1,24E+09 |
| Middle-earth_Men | 9 | 4 | 5 | 0.08333333333333333 | 2,20E+08 | 5,78E+08 |
| Middle-earth_Dwarves | 4 | 4 | 0 | 0.08333333333333333 | 0.0 | 1,02E+09 |
| Chemical_elements | 133 | 2 | 131 | 0.5 | 6,78E+09 | 1,82E+11 |
| Harry_Potter_characters | 13 | 12 | 1 | 0.0 | 5,06E+06 | 1,20E+09 |
| Ecology | 39 | 5 | 34 | 0.2 | 1,86E+11 | 1,83E+11 |
| Harry_Potter | 21 | 13 | 8 | 0.00641025641025641 | 5,43E+07 | 4,07E+09 |
| Babylon_5 | 18 | 7 | 11 | 0.047619047619047616 | 7,79E+07 | 8,35E+08 |
| Discworld_books | 6 | 6 | 0 | 0.0 | 0.0 | 1,02E+09 |
| Discworld | 17 | 7 | 10 | 0.023809523809523808 | 4,87E+06 | 5,01E+08 |
| Discworld_peoples | 2 | 2 | 0 | 0.0 | 0.0 | 1,02E+09 |
| Discworld_games | 3 | 2 | 1 | 0.0 | 1,85E+07 | 1,13E+09 |
| Games | 69 | 5 | 64 | 0.15 | 2,32E+10 | 0.00010877166918202868 |
| 1984 | 43 | 3 | 40 | 0.3333333333333333 | 4,32E+08 | 8,04E+09 |
| Animation | 35 | 5 | 30 | 0.15 | 1,94E+09 | 1,01E+11 |
| Doctor_Who | 26 | 10 | 16 | 0.03333333333333333 | 6,03E+08 | 4,64E+10 |

Bibliography

- [1] AGGARWAL, C. C., ZHAI, C., *Mining text data*. Springer Science & Business Media, 2012.
- [2] ANDRZEJEWSKI, D., ZHU, X., CRAVEN, M. “Incorporating domain knowledge into topic modeling via dirichlet forest priors”. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 25–32, New York, NY, USA, 2009.
- [3] BAEZA-YATES, R., RIBEIRO-NETO, B., OTHERS, *Modern information retrieval*. vol. 463. ACM press New York, 1999.
- [4] BEKKERMAN, R., ALLAN, J. Using bigrams in text categorization. Tech. rep., Technical Report IR-408, Center of Intelligent Information Retrieval, UMass Amherst, 2004.
- [5] BERRY, M. W., KOGAN, J., *Text mining: applications and theory*. John Wiley & Sons, 2010.
- [6] BIUK-AGHAI, R. P., NG, K. K. “A method for automated document classification using wikipedia-derived weighted keywords”. In: *2014 International Conference on Data and Software Engineering (ICODSE)*, pp. 1–6, Nov. 2014.
- [7] CAMPOS, D., MATOS, S., OLIVEIRA, J. L. “Biomedical named entity recognition: a survey of machine-learning tools”. In: , InTech, 2012.
- [8] CHAN, L. M., INTNER, S. S., WEIHS, J., *Guide to the Library of Congress classification*. ABC-CLIO, 2016.
- [9] CLAUSET, A., SHALIZI, C. R., NEWMAN, M. E. “Power-law distributions in empirical data”, *SIAM review* v. 51, n. 4, pp. 661–703, 2009.
- [10] ERDOS, P., RÉNYI, A. “On the evolution of random graphs”, *Publ. Math. Inst. Hung. Acad. Sci* v. 5, n. 1, pp. 17–60, 1960.
- [11] FELDMAN, R., SANGER, J., *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.
- [12] FELTOVICH, N. “Nonparametric tests of differences in medians: comparison of the wilcoxon–mann–whitney and robust rank-order tests”, *Experimental Economics* v. 6, n. 3, pp. 273–297, 2003.

- [13] FUCHS, C., BOERSMA, K., ALBRECHTSLUND, A., et al., *Internet and surveillance: The challenges of Web 2.0 and social media*. vol. 16. Routledge, 2013.
- [14] GABRILOVICH, E., MARKOVITCH, S. “Feature generation for text categorization using world knowledge”. In: *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pp. 1048–1053, San Francisco, CA, USA, 2005.
- [15] GABRILOVICH, E., MARKOVITCH, S. “Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge”. In: *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, pp. 1301–1306, 2006.
- [16] GABRILOVICH, E., MARKOVITCH, S. “Computing semantic relatedness using wikipedia-based explicit semantic analysis.”. In: *IJCAI*, pp. 1606–1611, 2007.
- [17] GADIRAJU, U., DEMARTINI, G., KAWASE, R., et al. “Human beyond the machine: Challenges and opportunities of microtask crowdsourcing”, *IEEE Intelligent Systems* v. 30, n. 4, pp. 81–85, July. 2015.
- [18] GADIRAJU, U., YANG, J., BOZZON, A. “Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing”. In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pp. 5–14, New York, NY, USA, 2017.
- [19] GANGEMI, A. “A comparison of knowledge extraction tools for the semantic web”. In: *Extended Semantic Web Conference*, pp. 351–366, 2013.
- [20] GANTNER, Z., SCHMIDT-THIEME, L. “Automatic content-based categorization of wikipedia articles”. In: *Proceedings of the 2009 Workshop on The People’s Web Meets NLP: Collaboratively Constructed Semantic Resources*, pp. 32–37, 2009.
- [21] GILES, J. “Internet encyclopaedias go head to head”, *Nature* v. 438, n. 7070, pp. 900–901, 2005.
- [22] GRISHMAN, R., SUNDHEIM, B. “Message understanding conference-6: A brief history”. In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.
- [23] HA, J., LEE, J., JANG, W., et al. “Toward robust classification using the open directory project”. In: *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 607–612, Oct. 2014.
- [24] HOFFART, J., YOSEF, M. A., BORDINO, I., et al. “Robust disambiguation of named entities in text”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 782–792, 2011.
- [25] HOVY, E., NAVIGLI, R., PONZETTO, S. P. “Collaboratively built semi-structured content and artificial intelligence: The story so far”, *Artificial Intelligence* v. 194pp. 2–27, 2013.
- [26] HU, J., FANG, L., CAO, Y., et al. “Enhancing text clustering by leveraging wikipedia semantics”. In: *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 179–186, New York, NY, USA, 2008.

- [27] HU, X., ZHANG, X., LU, C., et al. “Exploiting wikipedia as external knowledge for document clustering”. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 389–396, New York, NY, USA, 2009.
- [28] HUMPHRIES, M. D., GURNEY, K. “Network ‘small-world-ness’: a quantitative method for determining canonical network equivalence”, *PloS one* v. 3, n. 4, p. e0002051, 2008.
- [29] JOACHIMS, T. “Text categorization with support vector machines: Learning with many relevant features”. In: *European conference on machine learning*, pp. 137–142, 1998.
- [30] KITTUR, A., CHI, E. H., SUH, B. “What’s in wikipedia?: mapping topics and conflict using socially annotated category structure”. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1509–1512, 2009.
- [31] KITTUR, A., SUH, B., PENDLETON, B. A., et al. “He says, she says: conflict and coordination in wikipedia”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 453–462, 2007.
- [32] KRIZHANOVSKY, A. “Synonym search in wikipedia: Synarcher”, *arXiv preprint cs/0606097*, , 2006.
- [33] KUMAR, S., RENGARAJAN, P., ANNIE, A. X. “Wikitop: Using wikipedia category network to generate topic trees.”. In: *AAAI*, pp. 4951–4952, 2017.
- [34] LAN, M., TAN, C. L., SU, J., et al. “Supervised and traditional term weighting methods for automatic text categorization”, *IEEE Trans. Pattern Anal. Mach. Intell.* v. 31, n. 4, pp. 721–735, Apr. 2009.
- [35] LEIDNER, J. L., SINCLAIR, G., WEBBER, B. “Grounding spatial named entities for information extraction and question answering”. In: *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references-Volume 1*, pp. 31–38, 2003.
- [36] LÉVY, P., *Collective intelligence*. Plenum/Harper Collins New York, 1997.
- [37] LÉVY, P., *Cyberculture*. vol. 4. U of Minnesota Press, 2001.
- [38] MAKHOUL, J., KUBALA, F., SCHWARTZ, R., et al. “Performance measures for information extraction”. In: *Proceedings of DARPA broadcast news workshop*, pp. 249–252, 1999.
- [39] MANNING, C. D., RAGHAVAN, P., SCHÜTZE, H., *Introduction to Information Retrieval*. New York, NY, USA, Cambridge University Press, 2008.
- [40] MCCALLUM, A., NIGAM, K., OTHERS. “A comparison of event models for naive bayes text classification”. In: *AAAI-98 workshop on learning for text categorization*, pp. 41–48, 1998.
- [41] MCHUGH, M. L. “Interrater reliability: the kappa statistic”, *Biochemia medica: Biochemia medica* v. 22, n. 3, pp. 276–282, 2012.

- [42] MEADOW, C. T., BOYCE, B. R., KRAFT, D. H., *Text information retrieval systems*. vol. 20. Academic Press San Diego, CA, 1992.
- [43] MEDEIROS, J. F., NUNES, B. P., SIQUEIRA, S. W. M., et al. “Tagtheweb: Using wikipedia categories to automatically categorize resources on the web”. In: *European Semantic Web Conference*, pp. 153–157, 2018.
- [44] MEDELYAN, O., MILNE, D., LEGG, C., et al. “Mining meaning from wikipedia”, *International Journal of Human-Computer Studies* v. 67, n. 9, pp. 716–754, 2009.
- [45] MENDES, P. N., JAKOB, M., GARCIA-SILVA, A., et al. “Dbpedia spotlight: Shedding light on the web of documents”. In: *Proceedings of the 7th International Conference on Semantic Systems*, pp. 1–8, New York, NY, USA, 2011.
- [46] MIHALCEA, R. “Using wikipedia for automatic word sense disambiguation.”. In: *HLT-NAACL*, pp. 196–203, 2007.
- [47] MIHALCEA, R., RADEV, D., *Graph-based natural language processing and information retrieval*. Cambridge university press, 2011.
- [48] MILLER, G., *WordNet: An electronic lexical database*. MIT press, 1998.
- [49] MILNE, D. “Computing semantic relatedness using wikipedia link structure”. In: *Proceedings of the new zealand computer science research student conference*, 2007.
- [50] MITCHELL, J. S., BEALL, J., MATTHEWS, W., et al. “Dewey decimal classification”, *Encyclopedia of Library and Information Science*, , 1996.
- [51] NADEAU, D., SEKINE, S. “A survey of named entity recognition and classification”, *Linguisticae Investigationes* v. 30, n. 1, pp. 3–26, 2007.
- [52] NEWMAN, M., *Networks: an introduction*. Oxford university press, 2010.
- [53] O’REILLY, T., *What is web 2.0.* ” O’Reilly Media, Inc.”, 2009.
- [54] PETERS, I., *Folksonomies. Indexing and retrieval in Web 2.0*. Walter de Gruyter, 2009.
- [55] PONZETTO, S. P., STRUBE, M. “Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution”. In: *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 192–199, 2006.
- [56] RATINOV, L., ROTH, D., DOWNEY, D., et al. “Local and global algorithms for disambiguation to wikipedia”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pp. 1375–1384, Stroudsburg, PA, USA, 2011.
- [57] RIJSBERGEN, C. J. V., *Information Retrieval*. 2nd ed. Newton, MA, USA, Butterworth-Heinemann, 1979.
- [58] SALTON, G., BUCKLEY, C. “Term-weighting approaches in automatic text retrieval”, *Information processing & management* v. 24, n. 5, pp. 513–523, 1988.

- [59] SALTON, G., WONG, A., YANG, C. S. “A vector space model for automatic indexing”, *Commun. ACM* v. 18, n. 11, pp. 613–620, Nov. 1975.
- [60] SCHÖNHOFEN, P. “Identifying document topics using the wikipedia category network”, *Web Intelligence and Agent Systems: An International Journal* v. 7, n. 2, pp. 195–207, 2009.
- [61] SEBASTIANI, F. “Machine learning in automated text categorization”, *ACM Comput. Surv.* v. 34, n. 1, pp. 1–47, Mar. 2002.
- [62] SHAH, C., POMERANTZ, J. “Evaluating and predicting answer quality in community qa”. In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 411–418, New York, NY, USA, 2010.
- [63] SHIN, H., RYU, B.-G., RYU, W.-J., et al. “Bringing bag-of-phrases to odp-based text classification”. In: *2016 International Conference on Big Data and Smart Computing (BigComp)*, pp. 485–488, Jan. 2016.
- [64] SLATTERY, S., CRAVEN, M. “Combining statistical and relational methods for learning in hypertext domains”. In: *International Conference on Inductive Logic Programming*, pp. 38–52, 1998.
- [65] SOBHANA, N., MITRA, P., GHOSH, S. “Conditional random field based named entity recognition in geological text”, *International Journal of Computer Applications* v. 1, n. 3, pp. 143–147, 2010.
- [66] STEYVERS, M., TENENBAUM, J. B. “The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth”, *Cognitive science* v. 29, n. 1, pp. 41–78, 2005.
- [67] STRUBE, M., PONZETTO, S. P. “Wikirelate! computing semantic relatedness using wikipedia”. In: *AAAI*, pp. 1419–1424, 2006.
- [68] TAKEDA, M., KOBAYASHI, N., KITAGAWA, F., et al. “Classification of comments by tree kernels using the hierarchy of wikipedia for tree structures”. In: *Advanced Applied Informatics (IIAI-AAI), 2016 5th IIAI International Congress on*, pp. 123–127, 2016.
- [69] TAPSCOTT, D., WILLIAMS, A. D., *Wikinomics: How mass collaboration changes everything*. Penguin, 2008.
- [70] TURNER, V., GANTZ, J. F., REINSEL, D., et al. “The digital universe of opportunities: Rich data and the increasing value of the internet of things”, *IDC Analyze the Future*, , 2014.
- [71] VIÉGAS, F. B., WATTENBERG, M., DAVE, K. “Studying cooperation and conflict between authors with history flow visualizations”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 575–582, 2004.
- [72] VOSS, J. “Collaborative thesaurus tagging the wikipedia way”, *arXiv preprint cs/0604036*, , 2006.

- [73] WANG, P., DOMENICONI, C. “Building semantic kernels for text classification using wikipedia”. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 713–721, New York, NY, USA, 2008.
- [74] WANG, P., HU, J., ZENG, H.-J., et al. “Using wikipedia knowledge to improve text classification”, *Knowl. Inf. Syst.* v. 19, n. 3, pp. 265–281, May. 2009.
- [75] WATTS, D. J., STROGATZ, S. H. “Collective dynamics of ‘small-world’ networks”, *nature* v. 393, n. 6684, p. 440, 1998.
- [76] WILKINSON, D. M., HUBERMAN, B. A. “Assessing the value of cooperation in wikipedia”, *arXiv preprint cs/0702140*, , 2007.
- [77] YANG, Y. “An evaluation of statistical approaches to text categorization”, *Information retrieval* v. 1, n. 1-2, pp. 69–90, 1999.
- [78] YANG, Y., PEDERSEN, J. O. “A comparative study on feature selection in text categorization”. In: *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 412–420, San Francisco, CA, USA, 1997.
- [79] ZESCH, T., GUREVYCH, I. “Analysis of the wikipedia category graph for nlp applications”. In: *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, pp. 1–8, 2007.